Statistics in the Border Region

Sørensen, Nils Karl

Publication date:
2017

Citation for pulished version (APA):
Sørensen, N. K. (2017). Statistics in the Border Region: Compendium to statistics I and II BA-INT.

Nils Karl Sørensen ©
Statistics I and II, BA-INT
Edition 2017/2018

Statistics in the Border Region

Compendium to statistics I and II

BA-INT

1. Introduction and Topics

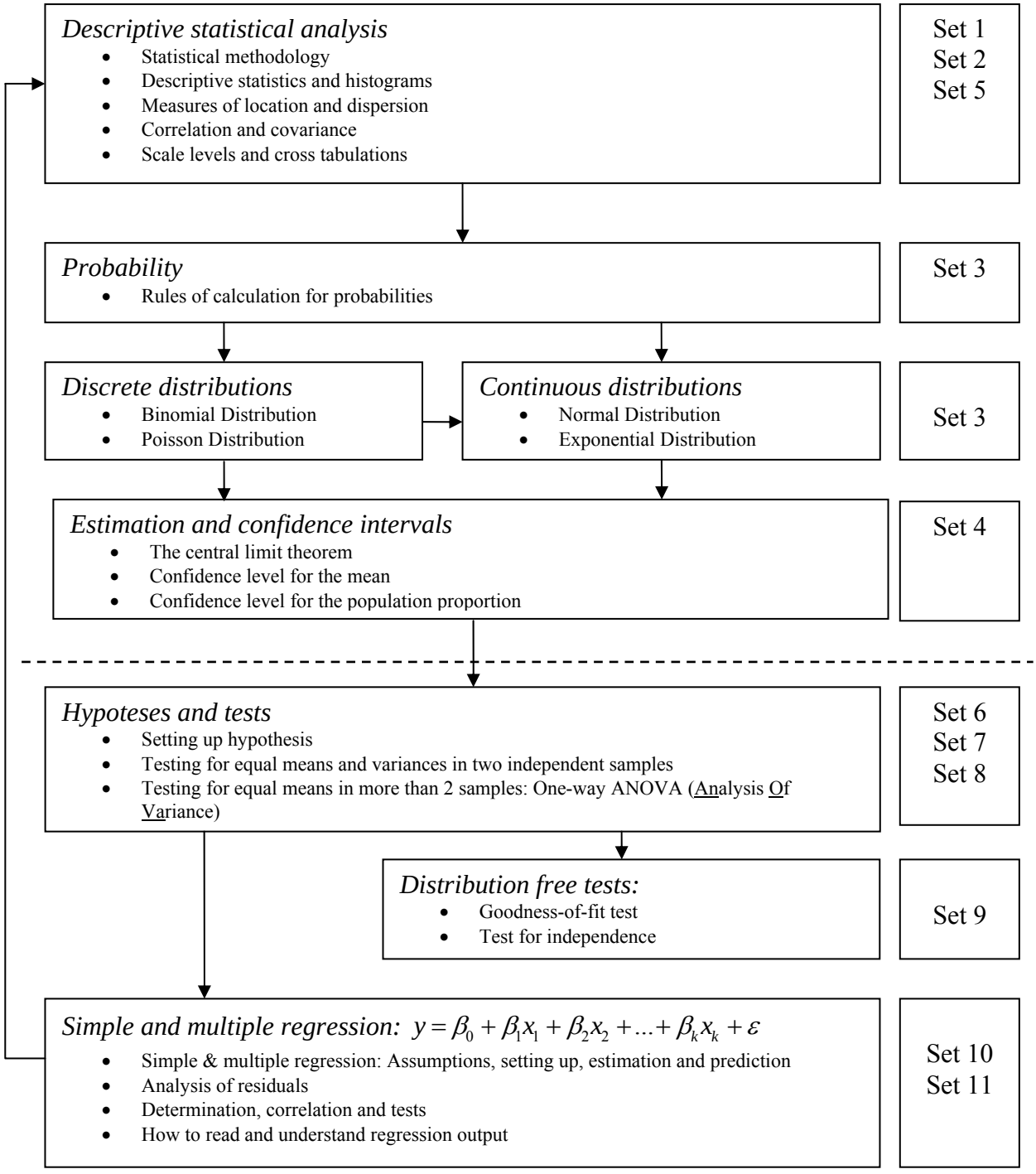
The present collection of notes constitutes the theoretical reference to the courses Statistics I and II at the BA-INT study. The course is build up as two separate series of lecturers and exercises each with a length of approximately 12 weeks. In addition to the notes there will be exercises to be solved separately. I had a good time developing this material. I wish you good and inspired studying ☺ Comments and improvements are always welcome!

Set	***Topics:***	***Pages***	***Total***
	Autumn Semester		***109***
1:	Statistical Methodology and Descriptive Statistics	39	
2:	Correlation and Covariance	5	
3:	Probability and Distributions	34	
4:	Estimation and Confidence Intervals	19	
5:	Scale Levels and Cross Tabulations	11	
	Spring Semester		***110***
6:	Setting Up Hypotheses and Simple Tests	16	
7:	Hypotheses and Tests in Two Independent Samples	26	
8:	One-way Analysis of Variance (ANOVA)	13	
9:	Goodness of fit Test and Test for Independence (χ^2-test)	13	
10:	Simple Regression Analysis	24	
11:	Multiple Regression Analysis	18	

There will be exercises to each set!

Outline of the Courses Statistics I and II

The upper part of the illustration above the dotted line shows the topics to be considered in Statistics I, whereas the lower part gives the topics from Statistics I. All topics are related to each other and form together a scientific synthesis. Roughly, a descriptive statistical analysis provides an initial description of data. Founded by the probability theory, hypotheses are stated and examined relative to data. Finally, the regression analysis gives a theoretical grounded analysis of relations among the variables considered.



Nils Karl Sørensen ©
Statistics I, BA-INT

Set 1: Statistical Methodology and Descriptive Statistics

by Nils Karl Sørensen

Outline

1. Statistical Methodology

My head is full of statistics because I have noticed that nothing can be proved without statistics

Mark Twain

In radio, television, Internet and on social networks, we are filled up with statistics everyday – from Eurostat, Eurosport, ARD, ZDF, DR to weather reports – in form of graphs, charts, tables etc. Data are reported in percentages, thousands, kilometres, squares etc. Statistical analyses are a necessity, when theory or problems related to theory needs to be confirmed or rejected. As quoted on the home page of Eurostat: *Statistics is a subject which can be applied to every aspect of our lives!*

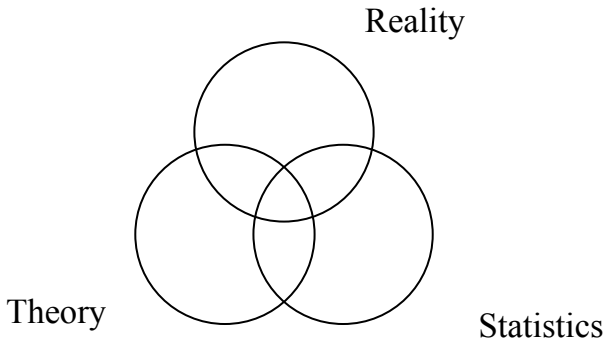
Statistics is numbers ("data"), and other material that can be assigned a value.

The word *statistics* origins from the Latin *statisticum collegium* (a public advisor) and the Italian word *statistica* (statesman or politician). The *statistica* was a person who compiled data for the state for example information regarding the composition of the population. This information was used for collecting taxes and military purposes. All information was – of course – a secret (No whistle-blowers)!

When an investigation is undertaken, we take our point of departure in the following three sets:

- Reality (as we observe it)
- Theory (based on our observations we set up theory)
- Statistics (based on our registrations of observations)

Consider the following diagram:



Good research is undertaken in the union (intersection) of the 3 sets. Notice that there are many possibilities of making errors. Observations of phenomena's will give rise to wonders, and this will result in formulations of theories. Statistics or another research method is a way to prove or reject the validity of the theories set up on our observations. So the use of statistics is an integrated part of an analytical work/research procedure.

The Danish astronomer Tycho Brahe (1546−1601) was one of the first scientists who saw this relation. I 1572 he observed a new and very light intense star on heaven (today this is called a Supernova). At that time it was the state of the art that heaven was a constant and static device. The observation led Tycho Brahe to the consideration that heaven was subject to change in time. In order to verify this hypothesis it was necessary to conduct systematic observations of heaven every night in order to identify changes. This led to constructions of statistics and tables and could serve for further analyses and investigations. In this way the project undertaken by Tycho Brahe is a direct application of the illustration above.

Statistics is used examine the interaction between theory and reality. In order for this to be valid the following must be possible:

1. The theory must be *operational.* That is it must be possible to set the theory into numbers.
2. It must be possible to find statistics that reflect the theory (verify or reject theory)

It is important to make clear that the interaction between theory and statistics implies *two sets of theory* each with a set of assumptions to be examined. There may be assumptions from the theory for example utility maximization from Microeconomics. The theories from statistics could be that the statistical method not is usable unless a given underlying statistical distribution is fulfilled. If this not is the case our results will be misleading or biased (to use a statistical term). Perhaps an example can clear things up a little.

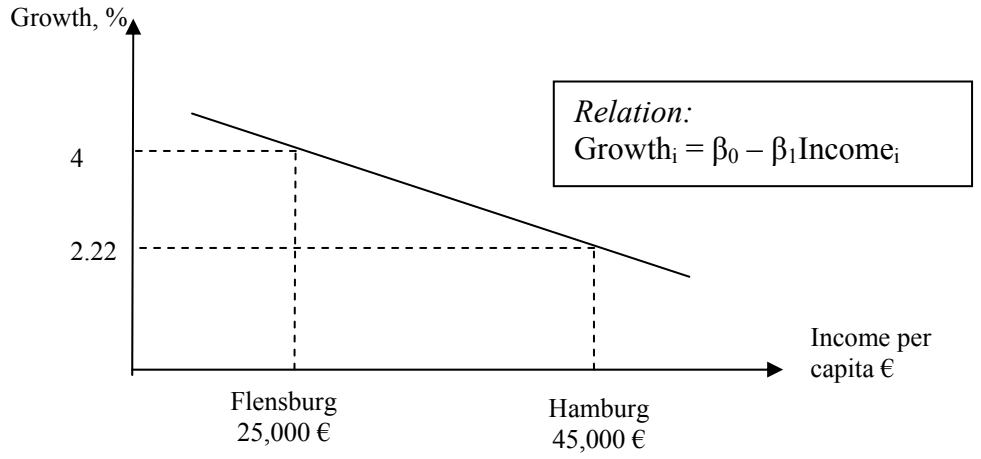
Discussion of a Model for Economic Convergence among Regions in Germany
Let us assume that we want to analyze the differences in regional income growth among German regions. As a result of differences in the composition of labor and capital from one region to another, differences will be observed in the rate of economic growth among regions.

For example we can use statistics from Eurostat or Statistisches Bundesamt on the regional income (GDP = Gross Domestic Product) per capita in Germany in a given year. For example, we can observe that the income per capita is higher in Hamburg than in Flensburg. This is due to the more central location of Hamburg and the larger harbor. However, the differences in income among the two cites also implies that it is easier to achieve a higher rate of growth in *Flensburg.* Why is this so?

Assume that the income in both cities is increased by 1,000 €. According to Eurostat, the income per capita in Hamburg equaled 45,000 € in 2004. Then the increase will be equal to

2.22 percent. However, in the same year, in Flensburg, the income per capita equaled 25,000 €. Here the increase equaled 4 percent. Consequently, because the level of income in Flensburg is lower than in Hamburg, the rate of growth in income per capita is higher.

Based on our calculations we can set up the following relation among economic growth and the level of income per capita:



We have plotted our two observations and connected them with a straight line. On these foundations, we can hypothesize a relation that states that the growth in less wealthy regions (like Flensburg) as a rule will be higher than in the more wealthy regions (like Hamburg). As a result in the long run there will be a "catch-up" among the region's leading to a more equal distribution of income. This is called the *economic theory of convergence*.

In the example we apply theories founded in other parts of economics. The model of convergence is for example founded in growth theory from Macroeconomics (VWL II). In order to make the model valid we apply statistical material.

Notice that in the illustration the straight line can in general be written as:

$$Growth_i = \beta_0 - \beta_1 Income_i \qquad i = 1,2, \ldots, n$$

This is the model that we are examining. It is consistent with economic theory. The line is valid for all *n* regions in Germany for example at the NUTS 2 level of regionalization. This is a division of Germany (and all other nations of the European Union) into a pre-specified level of regions. Further, *i* is the counting unit of regions for example a list of regions in alphabetic order, β_0 is the constant term or intercept, and the slope is equal to β_1. The model states that the slope is equal to the *rate of convergence*.

The two coefficients can be estimated by a mathematical model called *linear regression.* We shall return to this method in the course **Statistics II**, and we will actually use the method to estimate the rate of convergence for a given regional data set.

However, in order for regression to provide us with the correct estimates of β_0 and β_1 a number of assumptions have to be fulfilled for the *statistical model* to be valid. In the examined case the statistical material has to be normally distributed and independent. We shall return to the precise content of these terms later in set 3 of the notes to Statistics I. The check of assumptions of the model is called a *model control.* What if the assumptions not are fulfilled? Then the model is biased, and the estimates of the coefficients will be misleading. As a result the rate of convergence estimated will be wrong, and we do not even have a clue on nature of the bias i.e. if it is positive or negative.

The example has – hopefully – indicated that a statistical analysis is a complicated task to undertake. There are two types of statistics namely the *descriptive statistics* that we shall deal with in the course **Statistics I**, and the *inferential statistics* that are based on probability theory. We shall deal with this kind of statistics mostly in **Statistics II**. We can use the inferential statistical methods to see if the estimates are *significant* and to *test* if convergence is present.

How do we undertake a full Statistical Investigation?
The steps in a statistical investigation are as follows:

- Collection of material (e.g. taking a sample, using databanks)
- Presenting statistics (descriptive statistics)
- Preparation
    - Setting up hypothesis
    - Setting up a statistical model
    - Estimation of the model
    - Evaluation of the model and the underlying assumption
- Relevant conclusions to be drawn

The final part labeled *preparation* is also called ***statistical inference.*** Notice that a full investigation consists of a combination of *descriptive statistics* and *inferential statistics*.

Statistics and You
This is a very important issue! This is so because your statistics is not better than your presentations. Becoming a competent statistician is about more than learning the techniques outlined in these notes. You should be able to provide good and easy to read explanations of your findings. Many times I must say when I read exam exercises "What is going on" – I do not understand! Many central issues provided in reports can be illustrated by good and simple to read graphs, bar charts, tables etc. that can serve as point of departure of the more sophisticated analyses.

2. Using Excel and Pocket Calculator in Statistics

During the past decades courses in statistics has changed substantially. This is due to the emergence of computer and calculator technology. Menu driven programs such as spreadsheets implies that time consuming programming work has been minimized, and increased attention can now be given to the more analytical part of statistics. Today a course in statistics is more about setting up hypotheses, looking and interpreting computer output than before. This also implies that the exam to a higher degree is based problem related exercises.

In this section, I will deal with various types of means that is intended to *help* you in your work with statistics. ***Notice, that you are free to choose the software or calculator that you use.*** In many cases you can also use a so-called "**pocket calculator**" (need a large pocket!).

My own preference is the **Texas TI-84 calculator**. The **Texas TI-89 calculator** is more costly, advanced, but the screen is also more complicated to read for persons with myopia – and friendly speaking – I have myopia − and I can't see what is happening on the TI-89 screen. In addition, the interface is a little more complicated. You have to press F4 before you obtain exactly the same features as on the TI-84. Texas Instruments also supply a version of the TI-89 with a keyboard. It is called Voyage 200. It has the same features as the TI-89, and it also shares manual with this calculator.

Notice that the manuals has provided very good examples on how to use the calculator for statistics, and also when working with statistical distributions. I have provided the manuals in Blackboard in German, Danish as well as in English.

If you have an older calculator from HP or Texas then it is likely that it can perform most of the calculations. Much of this technology is quite old and developed during the late 1970ties and early 1980ties. My own Texas TI-30 from 1990 has for example similar features as the TI-30BX Multiview that is used in many public schools and gymnasiums today!

For the statistical distributions use my **Statistics Tables** also to be found in Blackboard. You can also install an app on your smartphone. For both Apple and Android such tables are available for free or at a low cost. The market changes constantly and I am not updated!

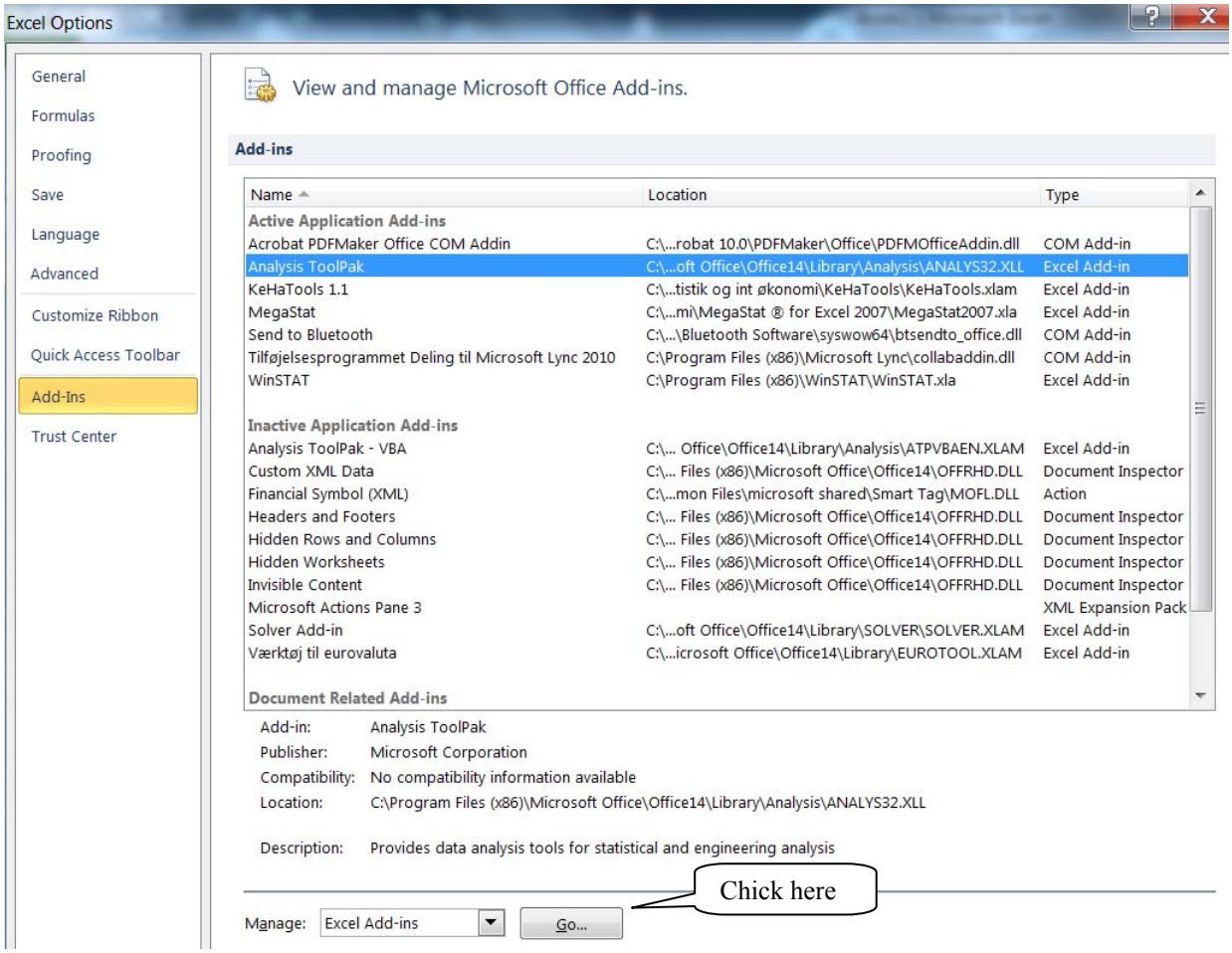
As just stated not so many years ago, a large part of a course in statistics was dealing with issues of programming! However, *all* the calculations needed in order to undertake a proper statistical investigation of a given dataset can be done by the spreadsheet **Excel**. In Excel we use the add-in ***Analysis Tool Pack***.

Analysis Tool Pack for Excel

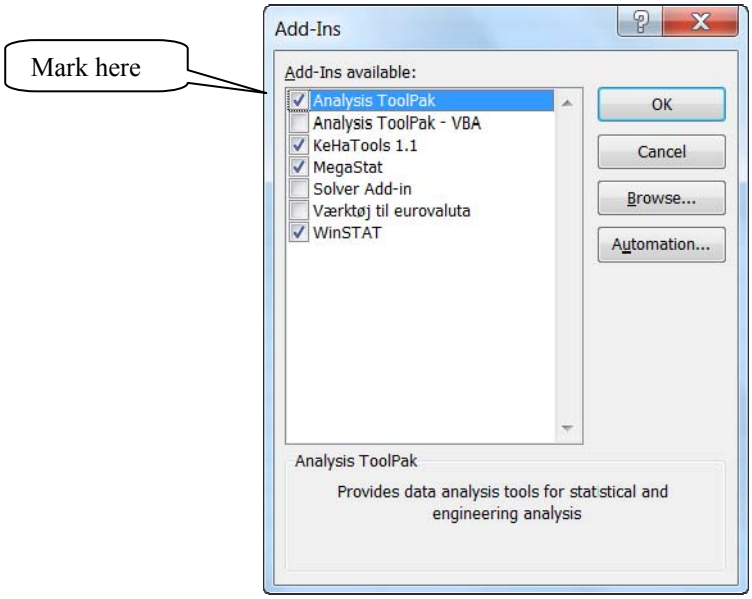
Below, I provide a short guide to set up your own computer with the *Analysis Tool Pack.* My outline is based on the Office 2010 package. Newer versions of the Office package have a slightly different design, but the methodology is the same.

Open first Excel → click on ***files*** → click on ***Excel options*** → click on ***add-ins***

The screenshot found on the next page should then appear. Here mark ***Analysis Tool Pack*** and click on ***GO***



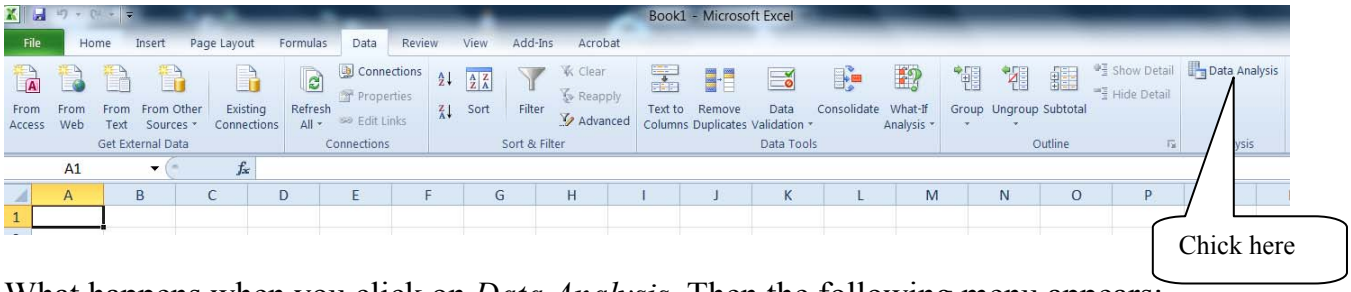
Now the following menu should appear:



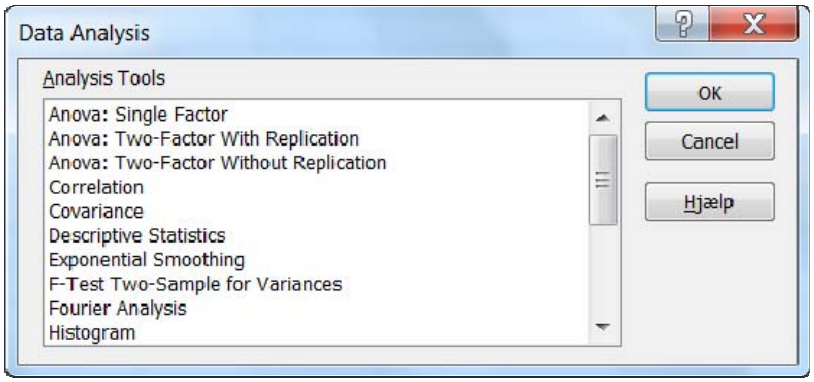
Mark *Analysis Tool Pack* and then ***OK***. Then you should be in the air for the statistical analyses ☺

How should the Toolbar look like?

Below is a picture of Excel toolbar with the menu *data* open. Here you find the *analysis Tool Pack* as *Data Analysis* to the right.



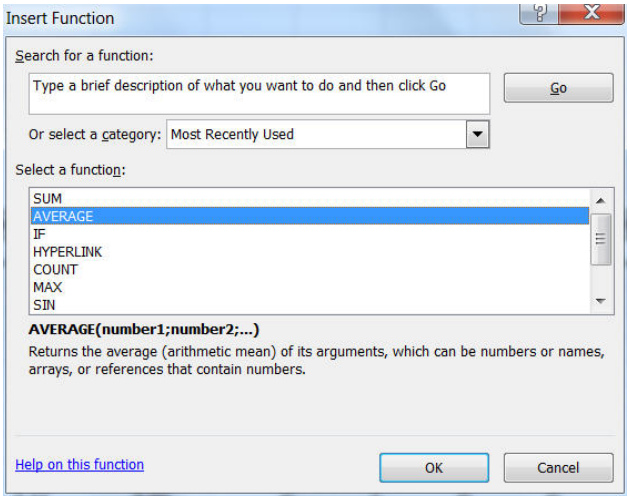
What happens when you click on *Data Analysis.* Then the following menu appears:



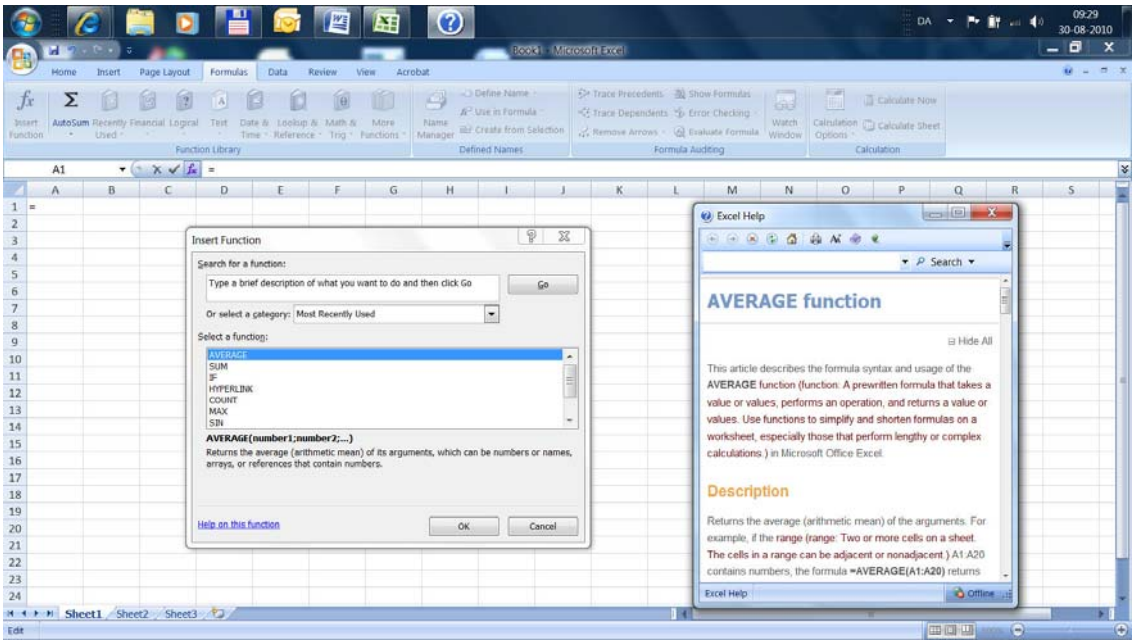
During the courses Statistics I and II you will be familiar with many of these functions!

Statistical Functions in Excel

Excel has another feature with relevance for these courses; namely ***function.*** Open Excel and click on ***Formulas.*** Then ***insert function*** is the first menu to the left. Click on this and obtain the screenshot:

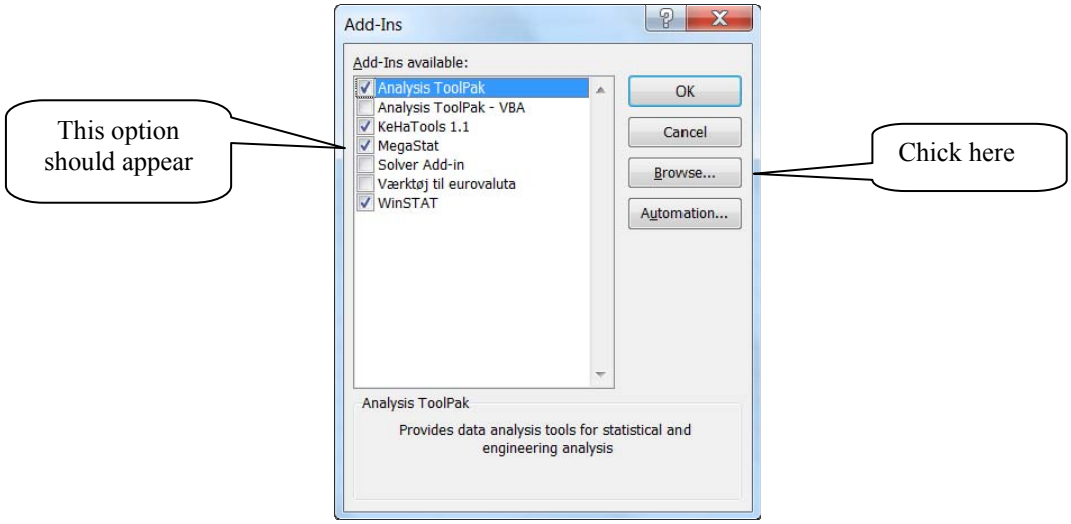


Having chosen ***insert function*** we get the menu above. We then select a category. The default is ***most recently used***. Let us instead select ***statistical***. Now we can select a specific function from the list organised by alphabetical order. Here, I have as an example, selected the function ***average.*** Notice, that below the function window a short description of the function is found. However, at more detailed description can be found by clicking ***help on this function.*** This information will be shown in the window to the left. This can be very helpful.

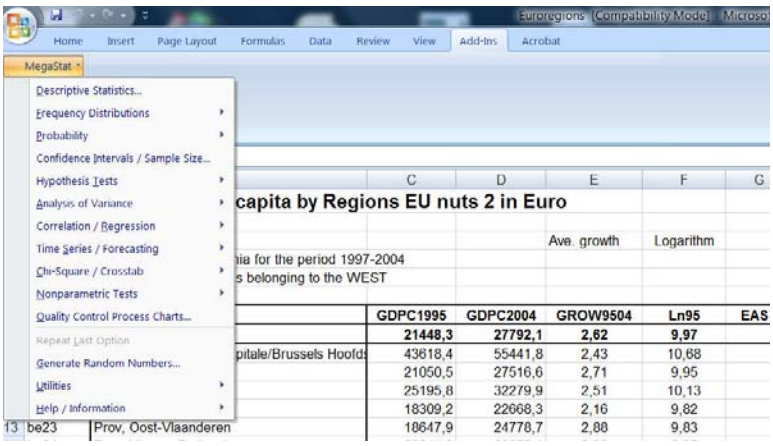


Megastat for Excel

Megastat is an Excel add-in that performs statistical analyses within an Excel workbook. The programme is installed by getting the file ***Megastat2007.xla*** from Blackboard under *Course Materials.* Locate the file in a folder called ABCD (or something). You can also find a manual in pdf-format for the add-in. After the file is downloaded on your PC open Excel and then under ***files → options → add-ins → GO.*** Now select ***Browse*** and find ***C:\ABCD\Mefastat2007.xla***. Then the word Megastat should appear in the Add-ins menu (as below). Mark Megastat, and then the program should be in the air.



After Megastat is installed, it appears on the Excel menu (Quick access toolbar in the top), and works like any other Excel option. Click on *add-ins*, and get the screenshot:



How to Buy Megastat etc.

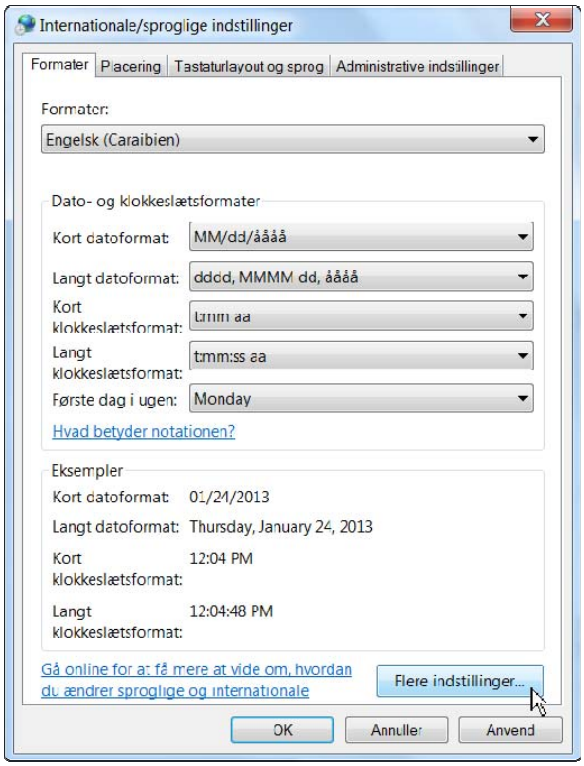
More recent versions may be available on the Internet. *Spurious bugs may also appear on the free download version.* It is possible to get a version of Megastat for MAC Office Package 2011 or later and also for Windows that works perfect. Consult the address: *MegaStat® for Microsoft® Excel® Windows® & Mac OS X®, Ver 10*

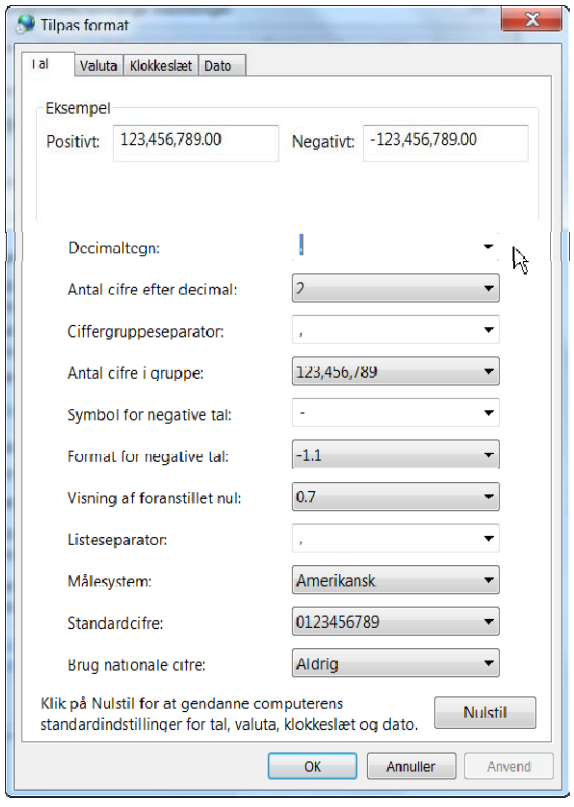
http://highered.mheducation.com/sites/0077425995/information_center_view0/index.html and continue under *first time user* and then *I am a student.* Follow the instructions and buy the official version. The price should be around 17 USD. **My suggestion is that you buy and install this version of Megastat.**

Megastat can perform statistical analysis similar to the *Analysis Tool Pack* in Excel (and sometimes even more). We will only use Megastat a few times for the present course where we use some options that are easier to work with than Excel. **If you work with the *Analysis Tool Pack* you do not need Megastat.**

A note for Scandinavian users only regarding the functionality of Megastat
In order for an old version of Megastat to run efficiently the comma separator has to be changed from a ”comma” to ”dot”.

This is undertaken in the ***control panel***. Here mark ***international settings/ laungage.*** Then the picture shown in the left panel below should appear. Here click on ***more settings (flere indstillinger).***





The menu to the right should now appear. Here change ***decimaltegn*** from “comma” to “dot” and also ***ciffergruppeseperator*** from “dot” to “comma”. Then click ***OK.***

(If this task not is undertaken Megastat may give you an error stating nonsense such as: ”probability must be between 0 and 1”).

3. Methodology of Descriptive Statistics

The purpose of conducting a descriptive statistical investigation for a single variable is to *clarify a number of characteristics of the examined data set*, i.e. its shape, variation, etc.

We want to undertake this investigation, so we can identify the underlying *theoretical statistical distribution* of the data generating process (DGP). For example the material can be Normal distributed or Uniform distributed. We shall return to this issue in the note set 3 to the present course. Here we shall deal with probability and statistical distributions.

If the data set has many observations it may be suitable to set up a frequency distribution. In such case we group the material into intervals. Frequencies can either be displayed in absolute or relative terms. If we use relative numbers the frequencies sum to 1 (or 100). So a frequency – just as the case with probabilities – ranges between 0 and 1. A *cumulative sum function* is a graph of the relative frequencies.

A descriptive statistical investigation consists of *three elements*:

- Graphical presentation for example by use of *a histogram*
- Presentation of measures of location = what is the typical?
- Presentation of measures of dispersion = how uncertain is the typical?

Let us first discuss how to provide a graphical presentation of the data set.

In statistics, we consider the following *types* of data:

- Cross-section: Many sectors/categories/regions at a given point in time
- Time series: One sector/category/regions over a period of time e.g. a year
- Panel: A combination of times series and cross section
- Census: Statistics provided through a questionnaire

Notice, that overlap among the types of statistics are present! In the example of the little regional model of convergence for the regions of Germany the data set is a cross-section data set. The data set is also a time series data set as growth is measured in time. Panel data sets are a little special. Here data for example are used to calculate the probability that a person will become unemployed by looking at how many time the person has been hit by unemployment earlier. These periods are referred to as spells of unemployment.

4. Graphical Presentations and Histograms

A graphical presentation is used in order to provide a good overview of the statistical material. Frequently, it will be necessary to count data. This task is undertaken be calculation of the *frequency*. The *frequency* is a measure of how often the item or observation is present in the data set.

A *histogram* displays classification into intervals of a quantitative variable. Normally the horizontal axis (x-axis) is the interval scale, whereas the vertical axis (y-axis) is used to display the frequency. The x-scale is often referred to reference point or the interval scale depending on the nature of the statistical material being considered. We shall return to this issue in the note set 5.

Let us as an example analyze a little data set with 20 observations of monthly incomes measured in 1,000 DKK. Data looks like:

9	6	12	10	13	15	16	14	14	16	17	16	24	21	22	18	19	18	20	17

The number of observations is equal to $n=20$.

This is a little bit unorganized, so let us rank the data by size from minimum to maximum:

6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24

Next interesting thing to observe is that there are several observations that take the same value. We can collect these observations and find the frequency. For example the value "16" can be observed 3 times, so the frequency is equal to 3. It is exactly this information we use when we set up a ***histogram***.

How can the data set be divided into some efficient categories or groups? I the following table the data set is attempted to be divided into intervals with a width equal to 5 by use of "arrows" or "marks" for each observation.

	Below 5	6 to 10	11 to 15	16 to 20	21 or more	Total
Number		|||	|||||	|||||||||	|||	20
Frequency	0	3	5	9	3	20
Relative %	0	0.15	0.25	0.45	0.15	1.00
Cumulative %	0	0.15	0.40	0.85	1.00	

Some additional calculations has been added to the table. The third line displays the *relative frequency* defined as $f_i = x_i/n$. Here x is the number of observations in the category i and n is the total. Finally, the fourth line displays the *cumulative frequency*. This is the summarized value of the relative frequencies, and the cumulative frequency sum to 1.00. Organizing data in this way we end up with an interval scale with 5 intervals.

How is the correct width found of a given interval? As a rule of thumb try out different sizes, and observe what the best is. A mathematical approach could be to apply the formula $\boldsymbol{2^k=n}$, where k is equal to the number of categories or intervals. As $2^4=16$ then we select $k=4$ categories as 16 is pretty close to 20. If $k=5$ we get 32, and it is too far away from 20.

The width of the interval can in general be found by use of the formula:

$$\frac{(x_{\max} - x_{\min})}{k} = \frac{24-6}{4} = 4.5$$

The intervals will now range as [6 to 10.5[; [10.5 to 15[; [15 to 19.5[and [19.5 to 24]

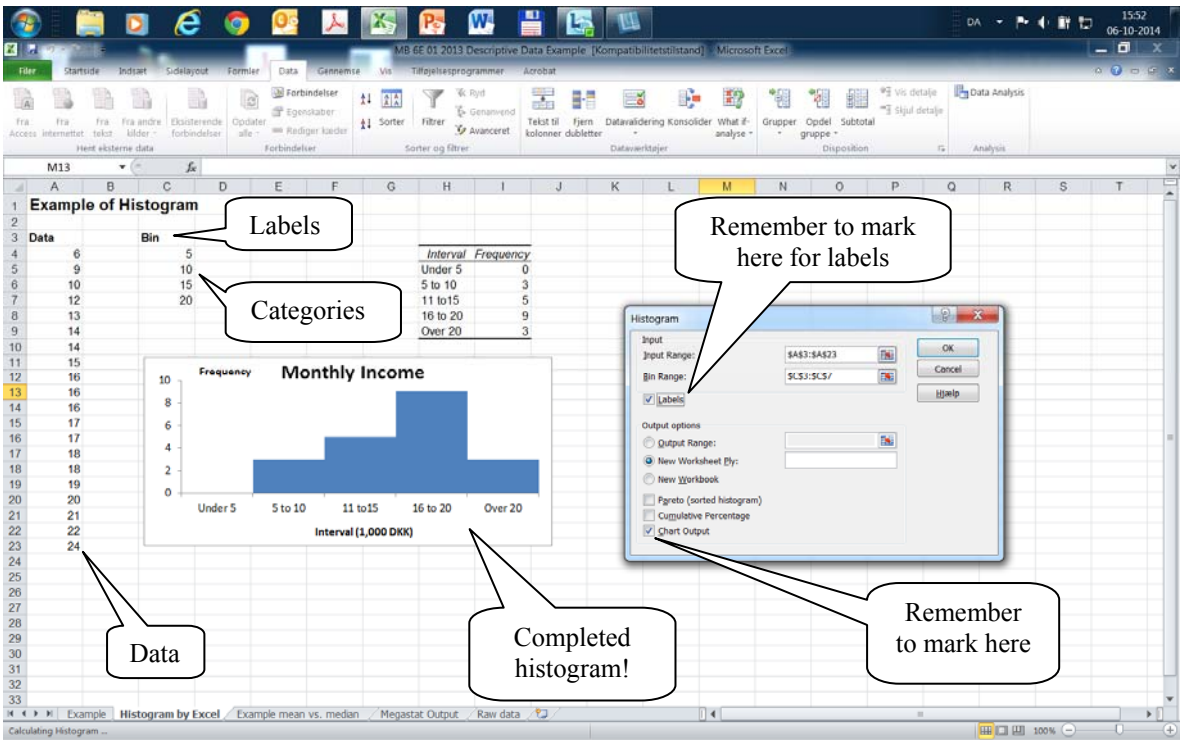
The table will look as:

	6 to 10.5	10.5 to 15	15 to 19.5	19.5 to 24	Total
Observations	||| ||||| |||||||| ||||	20
Frequency	3	5	8	4	20
Relative %	0.15	0.25	0.40	0.20	1.00
Cumulative %	0.15	0.40	0.80	1.00	

Independent of the applied approach the constructed histograms will be very equal.

Construction of a Histogram by use of Excel

Here click on ***data → data analysis → histogram.*** This task is undertaken in the screenshot below.



In the dialog box ***histogram*** you have to select or mark the data input area. Then the categories (called *bin range)* should be marked. The interval gives the lower bound. Remember to mark for labels (DK: Etiketter), and that chart output is required. In the final histogram in the screenshot a little bit of editing has been undertaken. Especially an attempt has been made in order to make the illustration as large as possible relative to all other text.

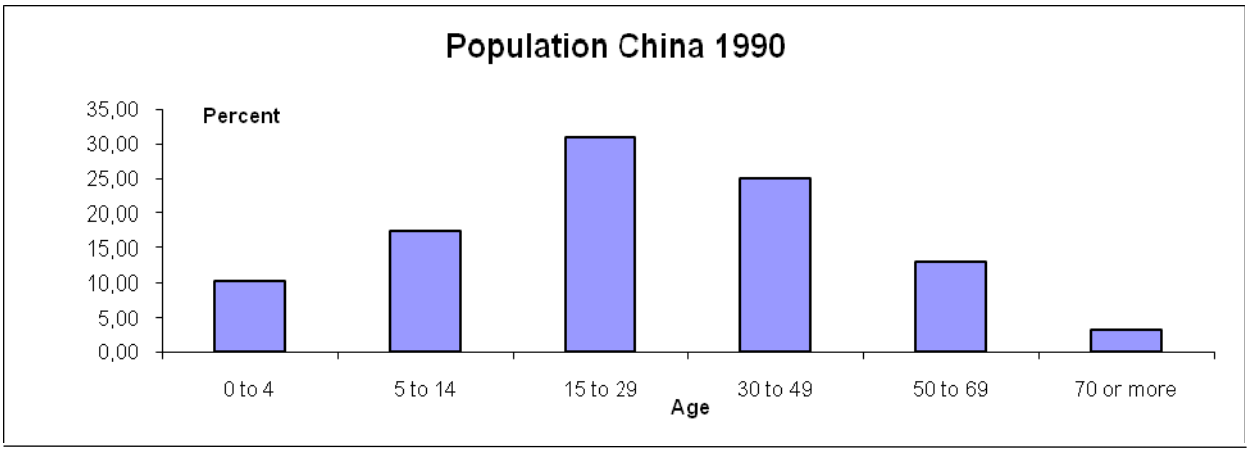
Histogram with Continuous Horizontal Axis and Varying Interval Width

Let us try to construct such a diagram by use of a different data set. Frequently, we run into the following problem of the "traditional" histograms" or "column charts" provided by Excel. Consider the following data set for the population of China July 1st 1990.

Age:	**0 to 4**	**5 to 14**	**15 to 29**	**30 to 49**	**50 to 69**	**70 or more**	**Total**
Persons, mill	116.60	196.90	350.50	283.10	147.90	36.80	1131.90
Persons, %	10.30	17.40	30.97	25.01	13.07	3.25	100.00
Units of 5 years	1	2	3	4	4	[4]*	18
% units of 5 years	10.30	8.70	10.32	6.25	3.27	0.81	

*=assumed

Using data from the first part of the table the following bar chart can be drawn:

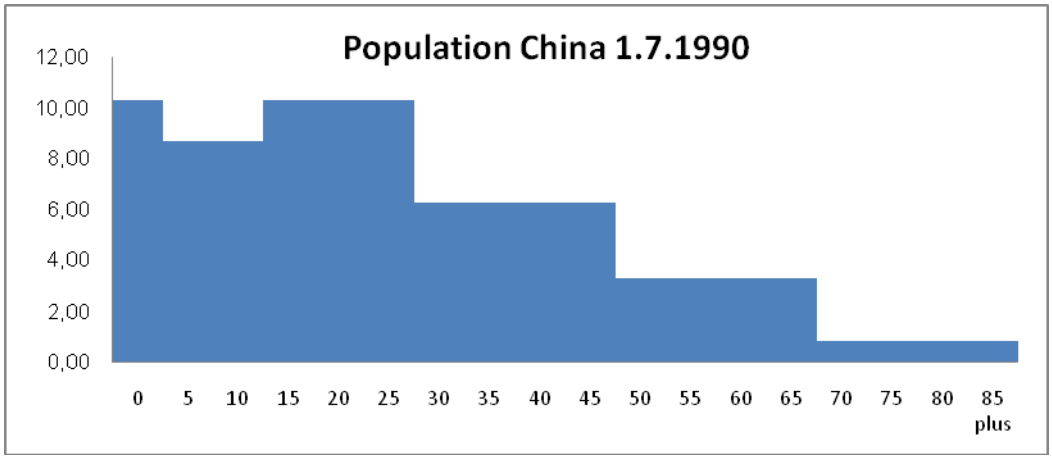


This is not a good bar chart, and it gives a misleading picture of data. *The graph is wrong because the intervals on the horizontal axis are not of equal length.* In order to solve this problem, we can for example divide data into units of 5 years. This task is undertaken in the lower part of the table above. For example the range from "5 to 14" contains two intervals of 5 years. So 17.40 is divided by 2 and 8.70 is obtained. The interval "70 or more" is a little special. Here the author has assumed a width of the equal to 4 units of 5 years. So the author assumes that a very limited number of Chinese in 1990 had an age exceeding 90 years. This judgment is set in the light of the very turbulent century in China with war, revolution etc.

To graph data we need to reorganize the spreadsheet a little. A table to use for the graph/histogram will look like the table on the top of the next page.

Age, year	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	>85
Person,%	10.3	8.7	8.7	10.3	10.3	10.3	6.25	6.25	6.25	6.25	3.3	3.3	3.3	3.3	0.8	0.8	0.8	0.8	0.8

In this case we use the normal bar chart function by Excel to draw the histogram. With a little bit of editing the histogram look as below:



Here the distribution of data is clearer, and more important; we can draw the correct conclusions from the graph. For example the most frequent categories are “0 – 5” and “15 – 29” years. Comparing with the first histogram it is observed that the first histogram overestimated the effect of the “one-child” policy undertaken in China from the 1970ties and on.

The issue of a data set grouped with a varying interval width appears quite often in for example population and labor market statistics. In the official Danish statistical ten year review a histogram on the distribution of agriculture farms by size of land having this problem has been published every year since 1976 although the histogram is incorrect and misleading.

5. Measures of Location

Measures of location or measures of position are an expression of the most frequent or typical observed unit. The reason for presenting different measures of location is that the different measures provides different amount of information. Some of these measures has earlier been presented in public school (7^{th} grade in Denmark) or in gymnasium.

We consider the following measures:

- Sample mean (arithmetic mean)
- Modus or mode
- Median
- Geometric mean
- Relation among the mean, mode and median
- Quartiles and percentiles

Let us just define the measures:

Mean, Sample Mean (Arithmetic Mean)
To calculate the mean all observations is used. The mean or arithmetic mean is the typical observation. For a total population[1] the mean is denoted as μ whereas the sample mean is denoted $\overline{X}$. The following formulas are valid:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For a *grouped dataset* with k groups or categories with the frequency f_i:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i \times x_i}{n}$$

Consider as an example the case with the data set the 20 observations of incomes used to draw the histogram in the previous section. The mean is found as:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6 + 9 + 10 + ... + 21 + 22 + 24}{20} = \frac{317}{20} = 15.85$$

[1] The notion between sample mean and total mean is important and we shall return to this issue later in note set 4.

In order to illustrate the calculations for a grouped data set consider an example of the distribution of grades for an exam in the course International Economics (VWL-III) that was held in February 2011 at the BA-INT study in Flensburg/Sønderborg. The distribution of grades on the Danish 7-point scale is given as

Grades of passed (7-point DK scale)	2	4	7	10	12	Total
Frequency	10	26	33	19	4	92

(I can inform that 8 students failed, so the share of passed students was fine) ☺

There are 7 categories or grades. The mean is calculated as:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i \times x_i}{n} = \frac{10 \times 2 + 26 \times 4 + 33 \times 7 + 19 \times 10 + 4 \times 12}{92} = \frac{593}{92} = 6.45$$

Sometimes when we have to work with grouped data where midpoints are used. We shall consider this issue in Section 8 of these notes.

Modus or Mode:
This is the most common observed observation, i.e. the observation with the highest frequency. In the example with the income data the mode is equal to 16, whereas the mode is equal to 7 (frequency equal to 33) in the grouped data set with the grades.

Median:
It is the middlemost observation when data are ordered. This is stated as:

Median = 0.50(n + 1) ordered position

In the data set with the incomes there are 20 observations so *n=20.* The median is then found at the ordered observation 0.50(20+1) = 10.5. This value is found in the table below giving the ordered distribution of the income data set. The median is equal to 16, and marked with yellow signature. The numbering of the observations is provided in the bottom line of the table.

Data	6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24
Frequency	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
Cumulative	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*

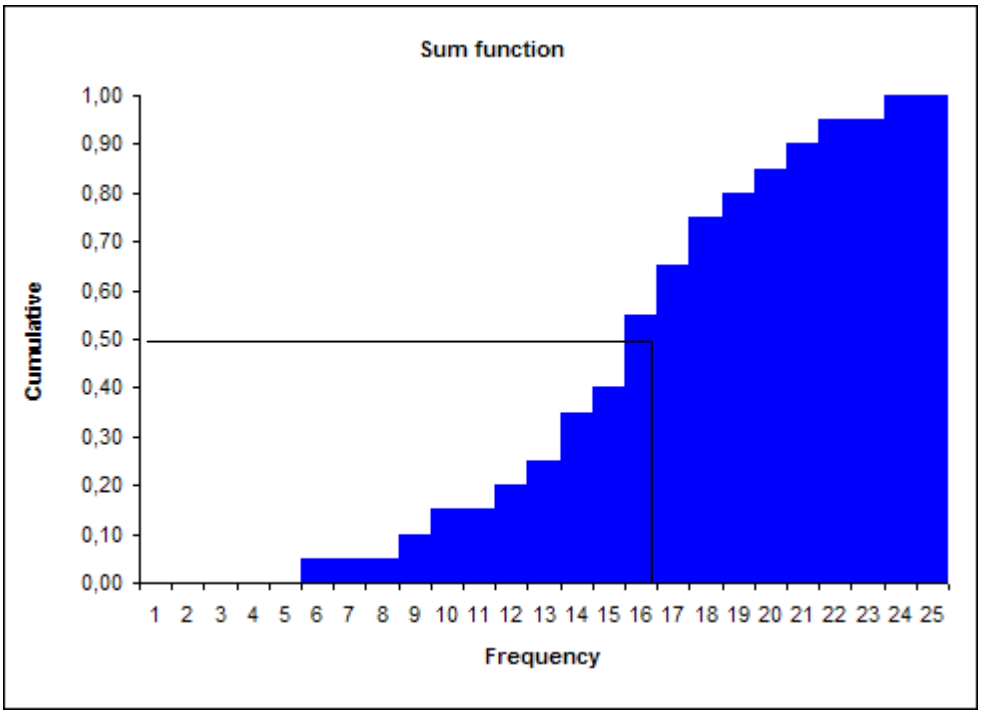
In the example with the distribution of the grades there are 92 observations. So the median is found at the 0.5(92+1) = 46.5 ordered observation. This observation is equal to the grade 7. Try to show this for yourself as an exercise!

Common for the mode and the median is that not all observations available are used for the calculation. The implication for the median is that extreme observations have no impact. This is very useful for example in wage negotiations and in wage statistics where it is the median wage that normally is displayed in the official statistics. Here the interest is on the typical wage. The wage of an extremely well-paid worker will influence the mean positively and result in a bias. This will not be the case if the median wage is used. In the final section of these notes this issue will be further elaborated.

The second and the third line in the table above bring the relative frequency defined as $1/n$ for each observation and the cumulative frequency summing to 100. The information in the first and the third line can be brought together to construct the *sum function* or *the cumulative frequency distribution*. The sum function is set up by use of a *simple bar chart* in Excel.

Sum Function by Diagram for Example

Obs	*Frequency*	*Cum. freq.*
1	0	0.00
2	0	0.00
3	0	0.00
4	0	0.00
5	0	0.00
6	1	0.05
7	0	0.05
8	0	0.05
9	1	0.10
10	1	0.15
11	0	0.15
12	1	0.20
13	1	0.25
14	2	0.35
15	1	0.40
16	3	0.55
17	2	0.65
18	2	0.75
19	1	0.80
20	1	0.85
21	1	0.90
22	1	0.95
23	0	0.95
24	1	1.00
25	0	1.00



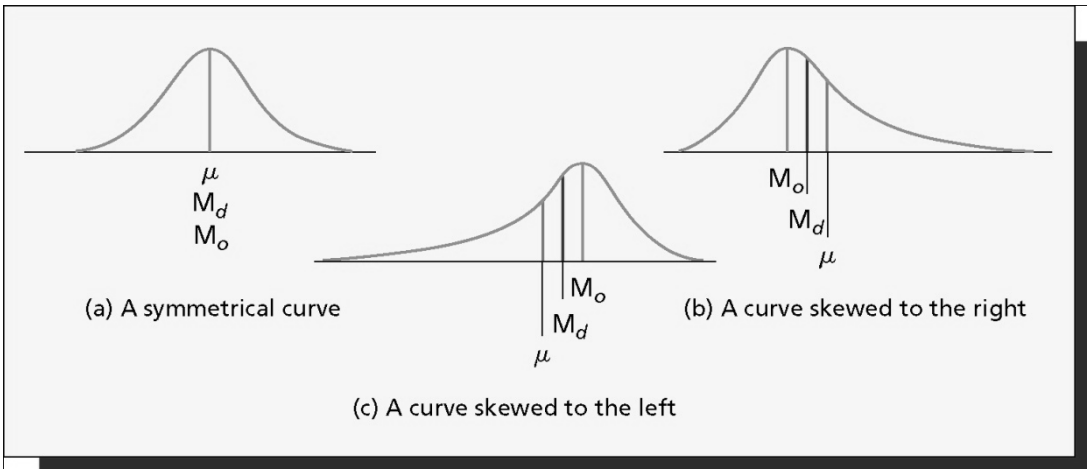
Notice a little bit of editing in Excel has been undertaken in order to obtain the sum function. Observe that the median has been inserted in the illustration.

The Shape of a Distribution of a Data Set, and the Relation among the Mean, the Mode and the Median

The illustration below brings together the relation among the mean $\boldsymbol{\mu}$, the mode $\mathbf{M_0}$ and the median $\mathbf{M_d}$. On the background of the relation among these measures of location, the following information can be deducted regarding the shape of the distribution of the examined data set:

a) Symmetry: $M_0 = M_d = \mu$
b) Skewed to the right: $M_0 < M_d < \mu$ (bulk of data left)
c) Skewed to the left: $\mu < M_d < M_0$ (bulk of data right)

It is observed that the median is the most robust measure of location whereas the mean is varying the most.



(M_0 = mode; M_d = median and μ = mean)

For the two examples considered it is found:

Income data set:

$\mu = 15.85 < M_0 = 16$ and $M_d = 16$ → data is weakly skewed to the left

Grade data set:

$\mu = 6.45 < M_o = 7$ and $M_d = 7$ → data is weakly skewed to the left

Finally, looking at the histogram for the distribution of the population for China on page 14 it is observed that the data set is skewed to the right.

Quartiles and percentiles

The formula to find the median can be generalized in order to find other positions in a given data set. Denote a *quartile* by Q_i. A quartile divides data into quarters and it is defined as:

$$Quartile = q(n+1) \text{ ordered position}$$

If $q = 0.25$ the *lower quartile* is observed. It is called Q_1. If $q = 075$ the *upper quartile* is observed. It is called Q_2. Finally if $q = 0.50$ the *median* is observed.

The *percentile* is defined as:

$$Percentile = p(n+1) \text{ ordered position}$$

Here p is a number ranging between 0 and 1. A *decile* is a special case. If $p = 0.1$ the *1st decile* is observed whereas if $p = 0.9$ the *9th decile* appear.

Let us use the formulas given calculate other *moments*. We can state the **5-point summary:**

1st decile	is 0.10-percentile	
Lower quartile	is 0.25-percentile	(called Q_1)
Median	is 0.50-percentile	
Upper quartile	is 0.75-percentile	(called Q_3)
9th decile	is 0.90-percentile	

Look again on the example with the 20 observations of incomes:

Data	6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24
Frequency	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
Cumulative	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*

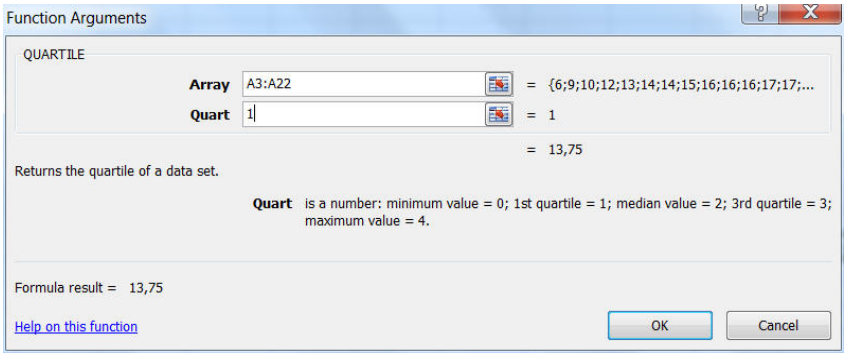
With our data listed above we obtain (we count the numbers of observations and look at the value (I obtained the exact values by use of Excel, see below)

10:	(20+1)(10/100) = 2.10	observations appears at	= 9.10
25:	(20+1)(25/100) = 5.25	observations appears at	= 13.75
50:	(20+1)(50/100) = 10.50	observations appears at	= 16.00
75:	(20+1)(75/100) = 15.75	observations appears at	= 18.25
90:	(20+1)(90/100) = 18.90	observations appears at	= 21.90

In **Excel** we find the *quartiles* by the function:

Formulas | ***Insert Function*** | ***Statistical*** | ***Quartile***

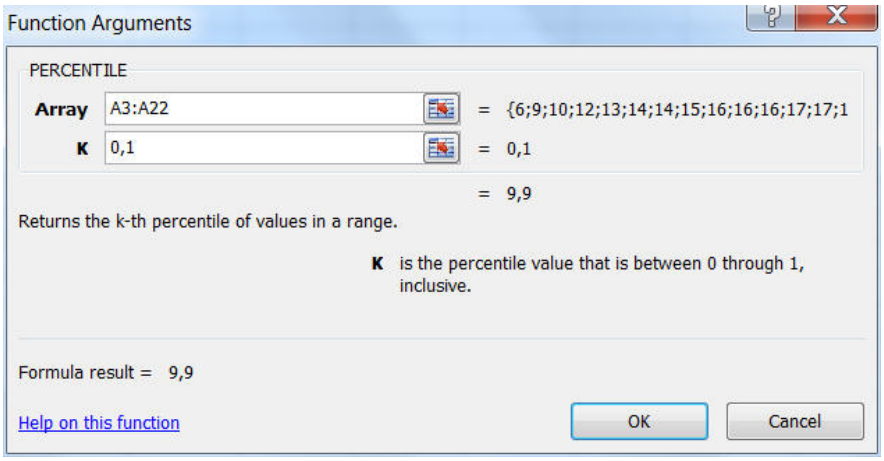
Press: 1 = Q_1, 2 = median, and 3 = Q_3.



In **Excel** generally we can find the *percentiles* by use of the function:

Formulas | ***Insert Function*** | ***Statistical*** | ***Percentile***

The value K is the percentile P ranging between 0 and 1.



In grouped observations we use interpolation, see Section 8 in these notes!

Geometric Mean:

Is defined as: $$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

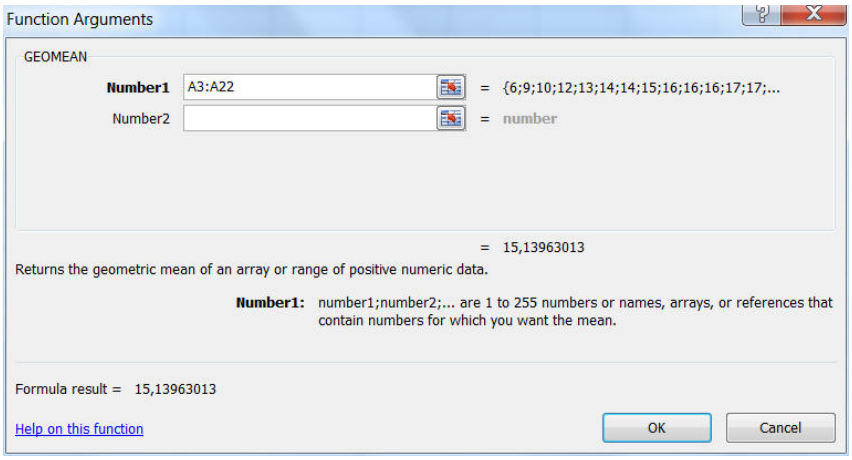
So the geometric mean is the *n* cubic root of the multiplicative sum of data. The geometric mean is always smaller than the arithmetic mean. This mean is commonly used in data set where the data generating process is assumed to be multiplicative. This is not very frequent in economics, and it is mostly used in relation to finance.

In the example with the income data the geometric mean is calculated as:

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[20]{6\times 9\times 10\times \ldots \times 21\times 22\times 24} = 15.14$$

In Excel: ***Formulas*** | ***Insert Function*** | ***Statistical*** | ***Geomean***

Then the following screenshot appears:



6. Measures of Dispersion and the Box-Plot

These measures describe the uncertainty around the typical observation.

We consider

- Range, inter quartile range, decile range and Box-plot
- Variance and standard deviation
- Coefficient of variation
- Skewness and kurtosis
- Chebyshev's inequality

Range, Inter Quartile Range, Decile Range and Box-plot
Common for all these measures of dispersion is they not use the all information in the data set.

Let us take our point of departure by considering the example on income above, and give the following definitions of the measures:

- The *range* is the distance between the largest and the smallest observation. For the income data set the range equals 24 − 6 = 18. For the data set with the grades the range is 12 – 2 = 10.
- The *inter quartile range (IQR)* is the difference between the upper and the lower quartile. We expect 50 % of the observations to be located within this range. For the income data set IQR equals 18.75 − 13.25 = 4.50
- The *decile range* is the difference between 9th and 1st decile. Here it equals 21.90 − 9.10 = 12.80. Within this range, 80 % of the observations are located.

A *Box-plot* is used for the identification of *outliers* and *suspected outliers.*

- An *outlier* can be defined as an observation more than 3 times the inter quartile range away from the first and third quartile.
- A *suspected outlier* is more than 1.5 (but less than 3) inter quartile ranges away from the first and third quartile.

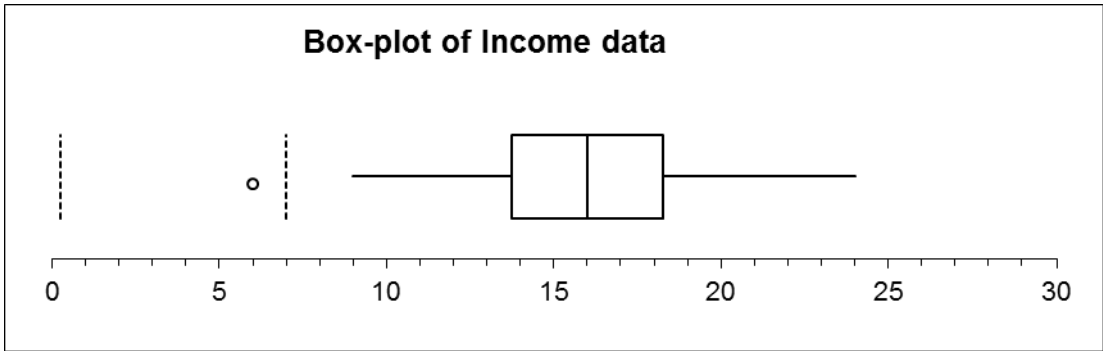
The Box-plot was invented in 1977 by John Tukey. A Box-plot can be criticized for having weak theoretical foundations for the use of the numbers "1.5" and "3.0".

However, the Box-plot is very suitable and popular for identifying extreme observations and in order to describe the distribution of a data set. We can calculate ”inner fence” and ”outer fence” as:

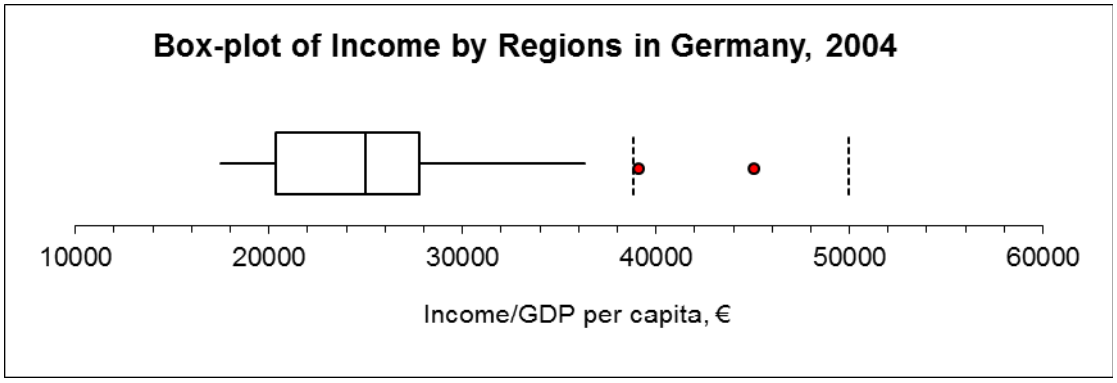
Lower inner fence: $Q_1 - 1.5\times IQR = 13.75 - 1.5(4.5) = \quad 7.00$
Lower outer fence: $Q_1 - 3.0\times IQR = 13.75 - 3.0(4.5) = \quad 0.25$

Upper inner fence: $Q_3 + 1.5\times IQR = 18.25 + 1.5(4.5) = \quad 25.50$
Upper outer fence: $Q_3 + 3.0\times IQR = 18.25 + 3.0(4.5) = \quad 32.25$

It is observed that the smallest observation (equal to 6) is a “suspected outlier”. We can draw the diagram manually by use of Word or as I have done hereby use of a program. For our data set the Box-plot look as:



Consider as another example the distribution of income per capita (GDP) by regions in Germany for 2004.



It is observed that 2 *suspected outliers* are present. It is Hamburg with an income per capita equal to 45,000 € and Munich (Oberbayern) with an income equal to 39,000 € per capita. The lowest level of income per capita is observed for the region Dessau.

Variance and Standard Deviation

These are the most used measures of dispersion. Contrary to the moments just listed these measures uses all the observations, and consequently uses all the information available in the data set.

The *variance* is the squared sum of all observations deviation from the mean divided by the number of observations. If we did not take the square the positive and negative deviations would cancel out. The variance easily becomes a large number as we are summarizing squared values. The *standard deviation* is the square root of the variance. This is a more traceable number.

The variance as well as the standard deviation can be calculated for the total population or the sample. The sample is smaller than the total population.

The *standard deviation* in a *sample* is given by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Here x_i is a given observation in the data set and n is the number of observations. We divide by $n-1$ in the sample because we have lost *one degree of freedom* at the time when the sample was taken. Compared to the total population the standard deviation in the sample will be a little larger. This is capturing the increased uncertainty in the sample because we not use all the observations that will be present in the total population.

The *standard deviation* for a *total population* is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

A frequently used formula for the standard deviation that is more complex to interpret, but good for calculus is given as:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}$$

For the data set on incomes the standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(6-15.85)^2 + (9-15.85)^2 + \ldots + (22-15.85)^2 + (24-15.85)^2}{20-1}} = 4.46$$

Using the more complex formula the standard deviation is found as:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]} = \sqrt{\frac{1}{20-1}\left[5{,}403 - \frac{(317)^2}{20}\right]} = \sqrt{\frac{1}{19}\left[5{,}403 - 5{,}024.45\right]} = \sqrt{19.92} = 4.46$$

Consider the standard deviation for a *grouped data set* with k categories and the frequency equal to f_i.

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}}$$

In the example with the grades data set the standard deviations is found as:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{10 \times (2-6.45)^2 + 26 \times (4-6.45)^2 + 33 \times (7-6.45)^2 + 19 \times (10-6.45)^2 + 4 \times (12-6.45)^2}{92-1}} = 2.83$$

Due to the large number of observations it can be debated whether this data set is a sample or a total population. If a division with n=92 is undertaken then σ = 2.81.

The Coefficient of Variation:

Gives the relative dispersion. It is much recommended for comparisons of different data sets. The coefficient of variation is equal to $CV = \frac{s}{\overline{X}}$ i.e. the standard deviation divided by the mean.

- If the distribution has large variation (is very flat) then CV takes a large value.
- If the distribution has small variation (is very steep) then CV takes a small value.

The coefficient of variation is also used when examining time series data for seasonal effect. If the seasonal variation is huge then CV is large

Skewness and Kurtosis

It is possible to calculate the degree of dispersion of a given data set away from symmetry.

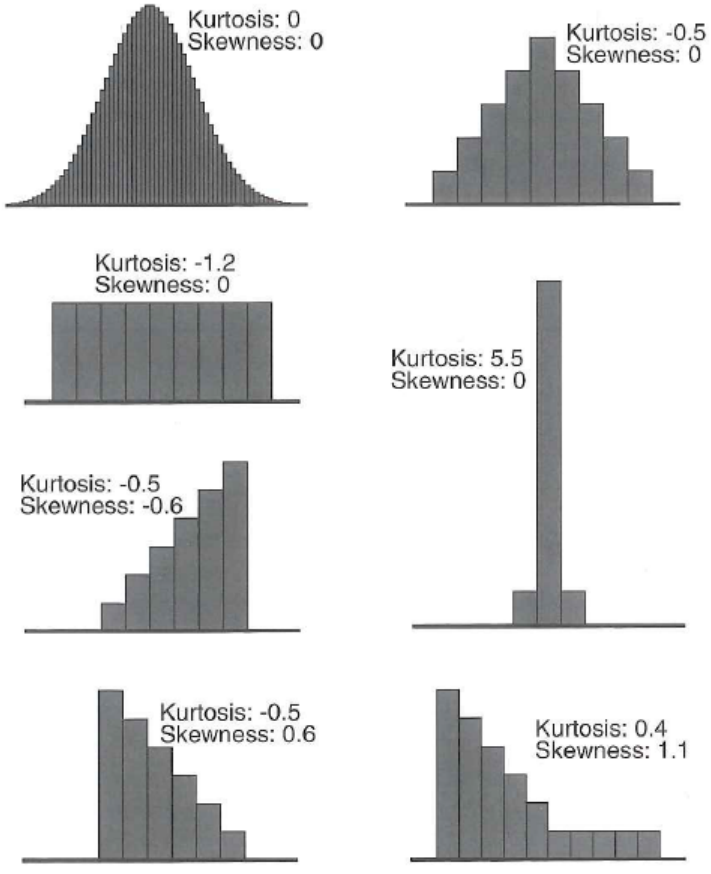
Skewness is an expression for how much the distribution is away from the ”normal”. If SK>0 data are skewed to the right, if SK=0 data are symmetric, and if SK<0 data are skewed to the left. The formula for skewness is given as:

$$SK = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 / n}{s^3} \qquad i = 1,2,\ldots,n$$

Kurtosis is a measure of the ”concentration” of the distribution”. If KU is large then we have a concentrated data set, and if KU is small we have a “flat” distribution. Kurtosis is given as:

$$KU = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 / n}{s^4} \qquad i = 1,2,\ldots,n$$

The illustrations below taken from E.M. Bøye, 2010, *Descriptive Statistics*, Swismark, page 134 gives some examples of values for SK and KU:



(Why can kurtosis be negative in computer programs like the Excel *Analysis Tool Package* and as shown in the illustration above when the formula reveals the data are powerd by 4? This is so, because many programs subtract –3 from *KU*).

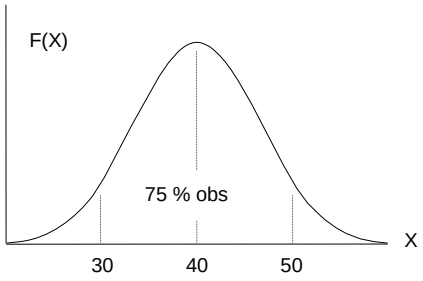
What is the interpretation of the numeric value of the kurtosis? If KU = 0, then we have exactly the normal distribution. This is also called *mesokurtic,* whereas KU > 0 is a *leptokurtic* distribution, and KU < 0 is a *platykurtic* distribution.

Chebyshev's Inequality
This is an empirical rule in order to make judgments on the variation of a distribution. Consider any population with mean μ and standard deviation σ. Then for any value *k>1,* at least $100(1-(1/k^2))\%$ of the population measurements lie in the interval [μ ±kσ].

Example
Let us assume that *k=2*. The we expect that at least $100(1-(1/2^2))\% = 100(3/4) = 75\ \%$ of the population observations will be found in the interval [μ ±2σ].

If the mean μ=40 and the standard deviation σ=5, then 75 % the observations will be found in the interval [40 ±2(5)] = [40±10] or [30 ; 50].

Illustration Chebyshev's Theorem:

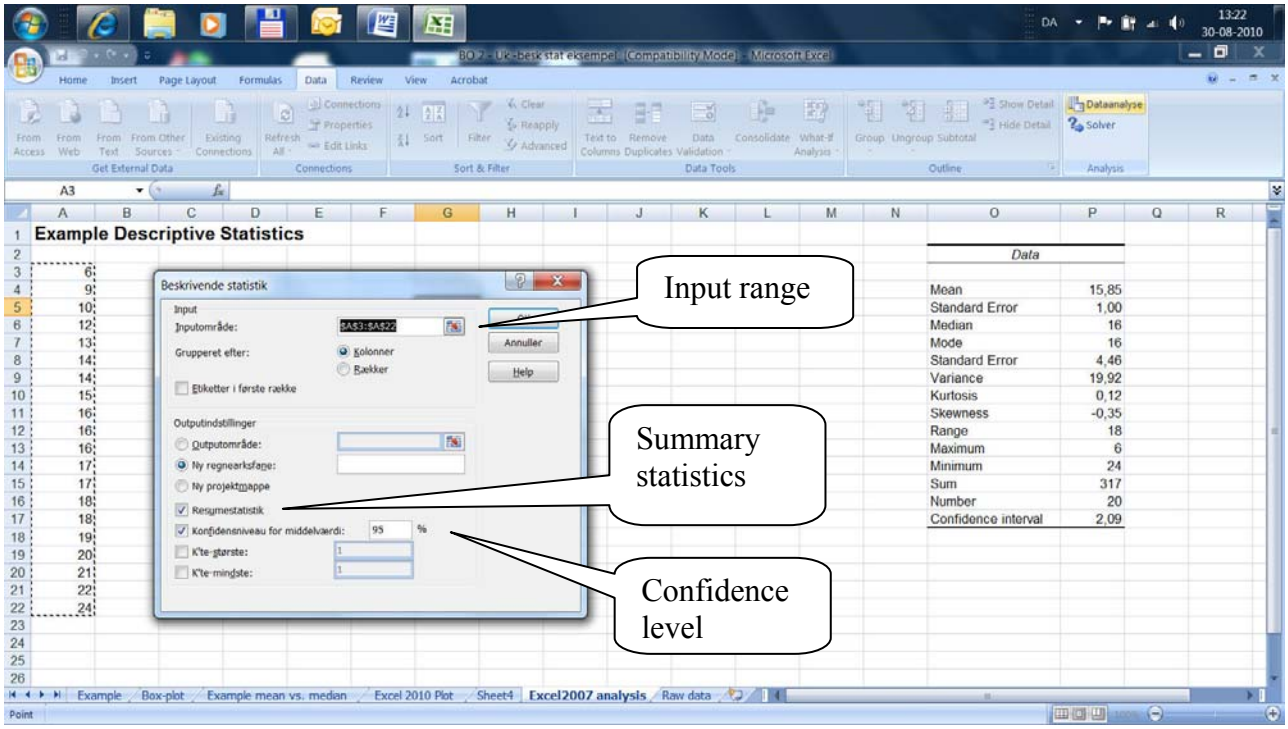


7. Descriptive Statistics on a Computer or Calculator

Descriptive Statistics by use of the Analysis Tool Pack in Excel
In order to obtain many of the measures just outlined undertake the following sequence:

1. Select ***data/data analysis/descriptive statistics.***
2. Mark data as input range.
3. Mark "labels" (if necessary), "summary statistics" and "confidence level for mean"
4. Click "ok".

Some highlights from the sequence is found on this screenshot.



Undertaking this procedure gives us the following output result for our little data set on income:

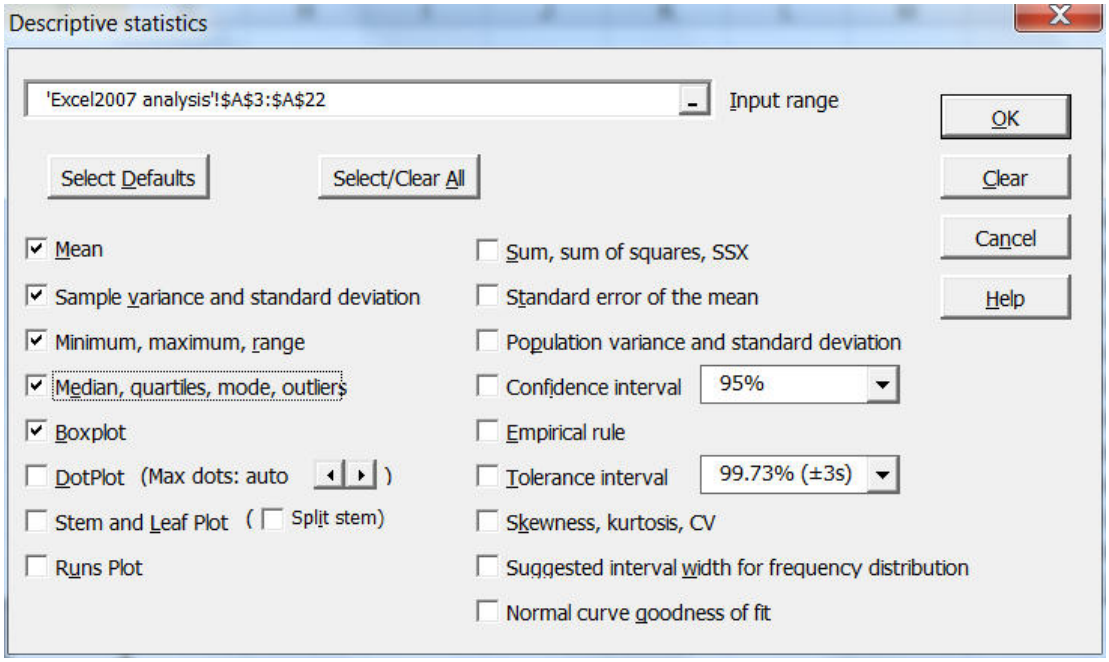
Income Data	
Mean	15.85
Standard Error	0.997
Median	16
Mode	16
Standard Deviation	4.46
Variance	19.92
Kurtosis	0.12
Skewness	-0.35
Range	18
Maximum	6
Minimum	24
Sum	317
Number	20
Confidence interval	2.09

Notice, that I have formatted data and only use 2 digits. Never use more digits than you need! We shall return to the confidence interval and the standard error defined as $SE = \frac{s}{\sqrt{n}} = \frac{4.46}{\sqrt{20}} = 0.997$ later in this course. The results in the table confirm many of the results that have been calculated in these notes.

Descriptive Statistics by use of Megastat

In order to draw a Box-plot by use of Megastat and select ***add-ins/Megastat/Descriptive Statistics***

Then we obtain the menu:



For our data we obtain:

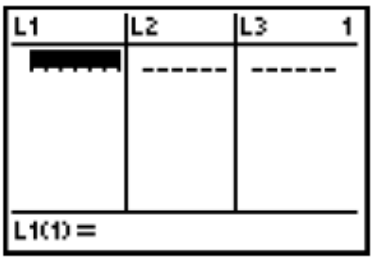
Descriptive statistics					
count	20	1st quartile	13.75	low extremes	0
mean	15.85	median	16.00	low outliers	1
sample variance	19.92	3rd quartile	18.25	high outliers	0
sample standard deviation	4.46	interquartile range	4.50	high extremes	0
minimum	6	mode	16.00		
maximum	24				
range	18				

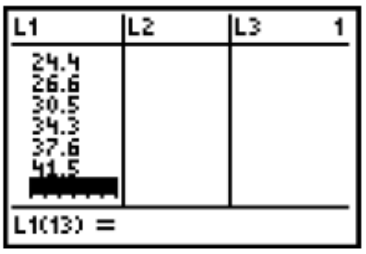
(The Box-plot has already been shown).

Descriptive Statistics by use of a Texas TI-84 Calculator

If you have a pocket calculator almost all the calculations presented above can be undertaken. It is a good idea to look in the manual. It can be found on the accompanying CD to the calculator or at the Internet at the address www.education.ti.com/ I have also uploaded the manuals in Danish, German and English on Blackboard.

Initially, on the **TI-84** data has to be loaded. There are 6 registers for data labeled L_1 to L_6. Pres STAT→1: EDIT→ENTER, the menu on the left next page will appear. Data are loaded as illustrated to the right.





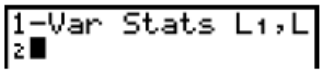
Many different manipulations can be done with the data – look in the manual se see the options!

Next step is to calculate the measures of location and dispersion. For a **normal data set** choose: STAT→CALC select now 1: **1-var stats** ENTER→**L_1**→ENTER. In this example the date set will be located in register L_1.

For a **grouped data set** as the data set on grades we have to load the variable (i.e. the grades 2,4,7,10 and 12) as well as the frequencies. Load the grades in the register L_1 and the frequencies in the register L_2.

Select on the calculator: STAT→CALC and then 1: **1-var stats** ENTER→**L_1,L_2**→ENTER

PS: "Comma" is found on the calculator just above "7 to 9".



A **box-plot** can also be set up by use of the calculator. If the data on income for example is loaded in the register L_1, then it is known that the minimum is equal to 6 and the maximum is equal to 24. So a width around 30 units is suitable for the plot. Select 2ND→STAT PLOT→1:ENTER (the key "STAT PLOT is found to the left just under the display). The plot function should be "ON" and the type of illustration is "Boxdiagram". Xlist is L_1 and now select WINDOW and set Xmin=0, Xmax=30. Finally, press ZOOM and select 9: ZoomStat ENTER. Then the box-plot should (hopefully) appear on the screen ☺

8. Descriptive Statistics in a Grouped Data Set

This section considers a more complex total population data set on incomes for Denmark 2013 with a division into intervals. The data set is divided by intervals and for each interval the mean income is given. If this not had been the case the midpoints should have been used instead.

How is the mean and standard deviation calculated? The formulas outlined earlier are used. The table below is provided for the calculations:

Disposal household incomes, Denmark, 2013

i	Interval for incomes 1,000 DKK	Number of households, 1,000 f_i	Mean income 1,000 DKK x_i	Income mass Mio. DKK $f_i \times x_i$	Deviation $(x_i - \mu)$	Square $(x_i - \mu)^2$	$f_i \times (x_i - \mu)^2$
1	0 - 49.9	413	8.2	3,387	-199.8	39,914.09	16,484,518
2	50 - 99.9	496	78.9	39,134	-129.1	16,662.96	8,264,830
3	100 - 149.9	752	127.1	95,579	-80.9	6,542.40	4,919,884
4	150 - 199.9	893	173.9	155,293	-34.1	1,161.79	1,037,482
5	200 - 249.9	760	224.2	170,392	16.2	262.92	199,821
6	250 - 299.9	553	273.1	151,024	65.1	4,239.95	2,344,692
7	300 - 399.9	499	340.1	169,710	132.1	17,454.35	8,709,719
8	400 -	284	643.0	182,612	435.0	189,237.96	53,743,582
Sum		4,650		967,131			95,704,528

Source: Statistics Denmark at www.dst.dk/ext/arbejde-loen-og-indkomst/rev13 attended 03-09-2015.

Mean and Standard Deviation

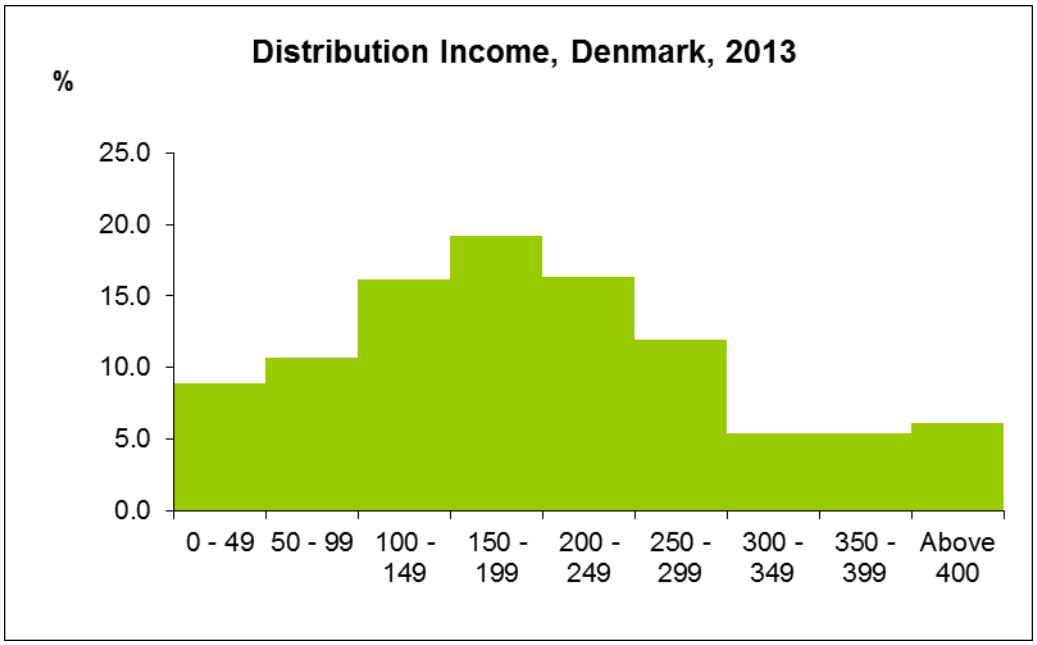
There are 8 categories i.e. $k = 8$. By insertion in the formulas:

Mean: $$\mu = \frac{\sum_{i=1}^{k} f_i \times x_i}{n} = \frac{967{,}131}{4{,}650} = 207{,}985\ DKK \approx 208$$

Standard deviation: $$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \times (x_i - \mu)^2}{n}} = \sqrt{\frac{95{,}704{,}528}{4{,}650}} = 143.46$$

There is some dispersion among the incomes. This is also evident from the histogram shown on the next page giving the distribution of the income by percent. For this calculation the table provided on the next page bottom serves as data source.

How can the median and the quartiles be calculated precisely in this grouped data set? For this task a procedure is applied called *linear interpolation.* This is a method where it within the interval is calculated the exact location of for example the median.



We take the point of departure in the table above. By use of the observations of the number of households a supplementary calculation can be undertaken in order to find the percentage distribution. This task is done in the table below. Initially, the relative frequency is calculated, and then the cumulative frequency.

Disposal household incomes, Denmark, 2013

i	Interval for incomes 1,000 DKK	Number of households, 1,000 f_i	Number of households frequency, % f_i/n	Cumulative frequency, %
1	0 - 49.9	413	8.9	8.9
2	50 - 99.9	496	10.7	19.5
3	100 - 149.9	752	16.2	35.7
4	150 - 199.9	893	19.2	54.9
5	200 - 249.9	760	16.3	71.3
6	250 - 299.9	553	11.9	83.1
7	300 - 399.9	499	10.8	93.9
8	400 -	284	6.1	100.0
Sum		4,650	100.0	

Source: Statistics Denmark at www.dst.dk/ext/arbejde-loen-og-indkomst/rev13 attended 03-09-2015

Quartiles, Median and Box-plot

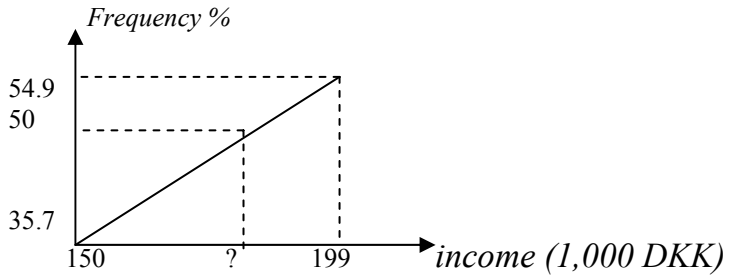
From the column with the cumulative frequency it can be observed that the lower quartile can be found within the interval of income equal to be between 100 and 149.9 thousand DKK. Similarly the median must be found within the interval between 150 and 199.9 thousand DKK. The upper quartile must be found within the interval ranging from 250 to 299.9 thousand DKK.

In order to obtain some more precise estimates *linear interpolation* is used. Within a given interval, we assume a uniform distribution of data. Then by using linear interpolation we adopt the following procedure:

We use a formula for example given as:

$$\textit{Value} = \text{"}\textit{End value interval}\text{"} - \frac{\text{"}\textit{too long relative to fractile}\text{"}}{\text{"}\textit{Total width in percent pct}\text{"}}\ \textit{interval width in value}$$

Illustration:



What happens in this illustration? The median is found at the middlemost observation at the percentile 0.5. From the illustration it is observed that at the percentile 54.9 an income is observed equal to 199.999 DKK. Further below at the percentile 35.7, incomes below 150,000 DKK are located. The median will be located within this interval.

Assume that the incomes are equally distributed over the interval. Then the income present at the percentile 0.50 can be found by subtraction of the share of incomes "too high" for this percentile. So we have to go 54.9 – 50.0 = 4.9 percentiles back relative to the total width of the interval. From the vertical axis the total width can be found as 54.9 – 35.7 = 19.2. This is also observed from the column labelled "number of households' frequency, %" in the table above.

The vertical axis is used for the calculation because here we have all information regarding the percentiles. This is not the case on the horizontal axis where we have to find the median income.

We now find the quartiles by linear interpolation, and application of the formula given above.

Median: $$200{,}000 - \frac{(54.9 - 50)}{19.2} \times 50{,}000 = 200{,}000 - 12{,}760 = 187{,}240 \qquad (M_d)$$

Similarly for the other quartiles and deciles:

Lower quartile: $$150{,}000 - \frac{(35.7 - 25)}{16.2} \times 50{,}000 = 116{,}975 \qquad (Q_1)$$

Upper quartile: $$300{,}000 - \frac{(83.1 - 75)}{11.9} \times 50{,}000 = 265{,}966 \qquad (Q_3)$$

Lower decile: $100{,}000 - \frac{(19.5 - 10)}{10.7} \times 50{,}000 = 55{,}607$

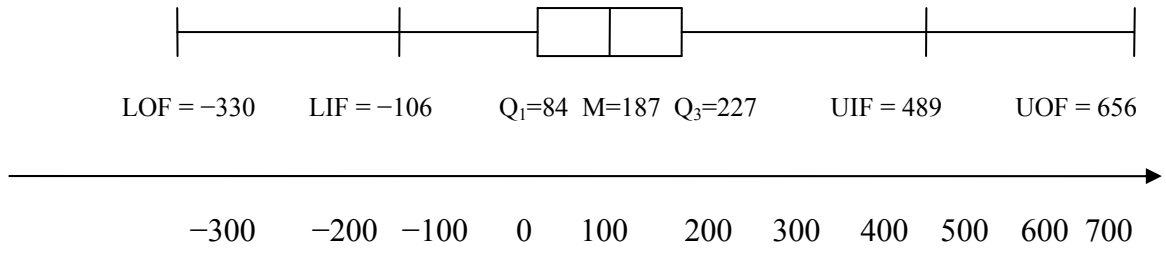
Upper decile: $400{,}000 - \frac{(93.9 - 90)}{10.8} \times 50{,}000 = 381{,}944$

Inter Quartile Range (IQR): $(Q_3 - Q_1) = 265{,}966 - 116{,}975 = 148{,}991$

As before we can set up a Box-plot:

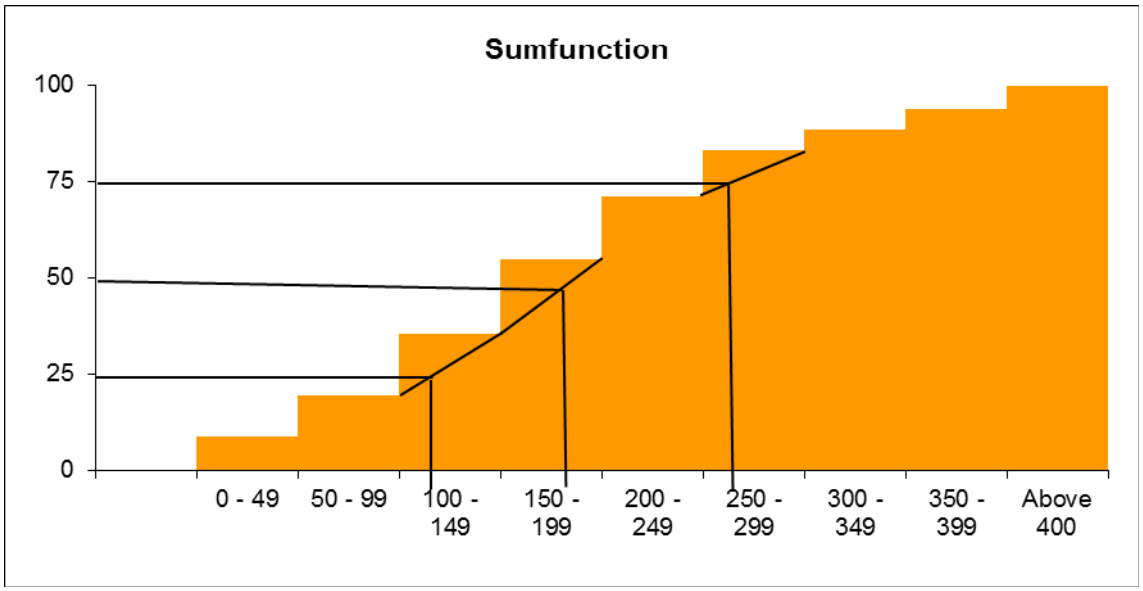
Lower inner fence:	$Q_1 - 1.5 \times IQR = 116{,}975 - 1.5(148{,}991) =$	−106,512
Lower outer fence:	$Q_1 - 3.0 \times IQR = 116{,}975 - 3.0(148{,}991) =$	−329,998
Upper inner fence:	$Q_3 + 1.5 \times IQR = 265{,}966 + 1.5(148{,}991) =$	489,452
Upper outer fence:	$Q_3 + 3.0 \times IQR = 265{,}966 + 3.0(148{,}991) =$	656,476

The Box-plot for this grouped data set can be drawn as follows:



The mean was found to equal 208 whereas the median is equal to 187. Finally, the mode interval is from 150−199. As $\mu > M_d$ the distribution is (weakly) skewed to the right.

Next, we show the sum function (here I have also added lines indication the location of the quartiles). The information for this illustration comes from the fourth column in the table on page 34. I have inserted lines for the lower quartile, the median and the upper quartile at the vertical axis. Assuming uniformly distributed incomes within each interval I can find the associated values on the horizontal axis. It can be observed that the findings are quite consistent the calculations performed above.



In an earlier version of the notes I used statistics on income for year 1987. Finally, a comparison of the distribution of the disposal income can be given as:

	Mean	Standard deviation	Lower Decile	Lower Quartile	Median	Upper Quartile	Upper Decile	Mode
1987	165,648	104,730	56,343	84,328	144,149	227,365	293,684	50,000−99,999
2013	207,985	143.460	55,607	116,975	187,240	265,966	381,944	150,000−199,999

It is observed that the mean disposal income increased by 25.5 percent during this 26 years period. This is equal to 0.88 percent annually. Observe that the lower decile has decreased whereas all other indicators have increased. The increase in the upper decile is equal to 1.02 percent annually. The conclusion is that incomes have become more unequal distributed.

How do we calculate the growth rates? For example for the mean disposal income use the formula:

$$\overline{G}_{n..0} = \sqrt[n]{\frac{x_n}{x_0}} - 1 = \sqrt[26]{\frac{207{,}985}{165{,}648}} - 1 = 0.00879$$

Then multiplying by 100 and obtain 0.879 percent. The period n is equal to 2013−1987 = 26.

9. Descriptive Statistics – an Example of Outliers

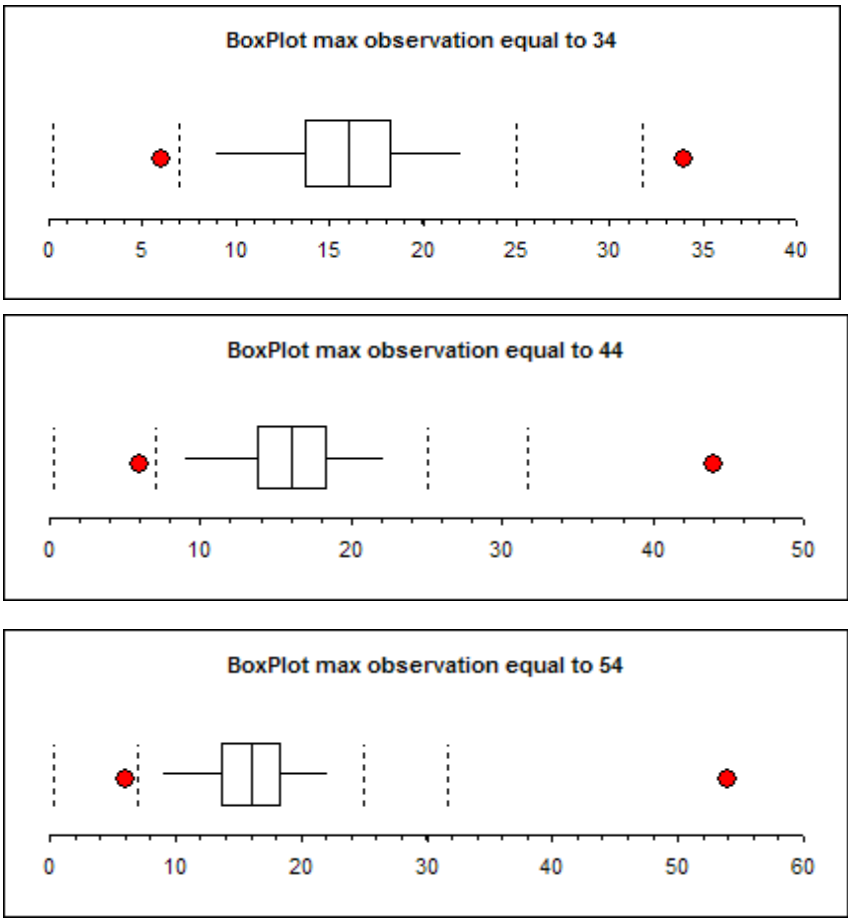
Let us finally conduct a little experiment to examine what happens with the mean and the median in the case where we consider extreme observations or outliers.

As summarized earlier for our little data set:

	Descriptive statistics		
Mean		15,85	Smaller than median and mode
Geometric Mean		15,14	This is smaller
Standard Error		0,998	s/(root n)
Median		16	
Modus / Mode		16	
Standard deviation		4,46	
Sample variance		19,92	
Kurtosis		0,12	A little steep distribution
Skewness		-0,35	Skewed to the left
Range		18	
Minimum		6	
Maximum		24	
Sum		317	
Observations		20	
Confidence interval(95,0%)		2,09	

In order to explore the stability of the median, I have replaced the observation "24" by "34", "44" and "54" respectively. On the next page see the descriptive statistics and the accompanying box-plots:

	Basic	*Max=34*	*Max=44*	*Max=54*	
Mean	15.85	16.35	16.85	17.35	Increases
Standard Error	1.00	1.29	1,69	2.13	
Median	16	16	16	16	Constant!!
Modus / Mode	16	16	16	16	
Standard deviation	4.46	5.79	7.56	9.52	
Sample variance	19.92	33.50	57.08	90.66	
Kurtosis	0.12	3.88	8.99	12.55	
Skewness	-0.35	1.19	2.43	3.16	Increases
Range	18	28	38	48	
Minimum	6	6	6	6	
Maximum	24	34	44	54	
Sum	317	327	337	347	
Observations	20	20	20	20	
Confidence interval(95 %)	2.09	2.71	3.54	4.46	Increases



It is observed that we (of course) get an extra outlier! Notice that the median is constant, whereas the mean and variance (as well as the confidence interval) increases. Further, Skewness gets more positive indicating a distribution skewed to the right i.e. $\mu > M_d$.

Nils Karl Sørensen ©
Statistics I, BA-INT

Set 2: Correlation and Covariance

by Nils Karl Sørensen

Outline ***page***

1. Correlation and Covariance

In the former set of notes on *descriptive statistics* we examined the relation among two variables in the example on *convergence.* The considered variables were the rate of economic growth and the level of income.

We can label the two variables x and y. By setting up a scatter chart for n pairs of observations (x_i,y_i) for all observations $i = 1,2, \ldots n$ we can obtain information with regard to the strength of the relationship among the two variables. Such a relation and be positive, negative or non-existing.

The *covariance* is an indicator for the strength of the relation. The *covariance* is labelled S_{xy} and can be defined as:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

In order to calculate the covariance we first have to find the means for x and y. Next for each pair of observations the difference to the mean is calculated, and the outcome is multiplied. Finally, data are summarized and divided by $(n-1)$.

The covariance can be positive, negative or zero. In the last case no relation is observed among the two variables.

A problem with the covariance is that it depends on the measurement unit of x and y. If we for example measure in Danish currency DKK then the covariance will be numeric much smaller than in the case where the measurement unit is equal to 1,000 DKK.

This problem can be solved by undertaking a "normalization" of the covariance. In such a situation a factor is introduced such that the covariance becomes much smaller, and in addition, it is possible to compare with the covariance of other datasets.

The *correlation coefficient* is an expression for the normalized strength of the covariation among two variables x and y. For the normalization the sample standard deviations are used. The *correlation coefficient* can be defined as:

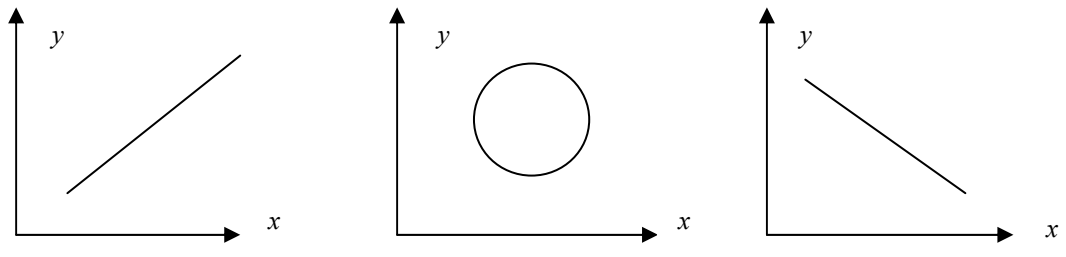
$$r = \frac{S_{xy}}{S_x S_y}$$

Here S_x and S_y denote the sample standard deviations for the data sets x and y respectively. The sample standard deviations are defined as

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{and} \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Using this normalization implies that the correlation coefficient will range between $-1 \le r \le 1$. The closer the coefficient of correlation approach to ± 1 the stronger is the correlation to be observed. If the coefficient of correlation approaches zero, then no relation is observed among x and y.

Examples of Correlation Coefficients in Scatter Diagrams



Positive **No relation** **Negative**

The correlation coefficient can be used to calculate the *coefficient of determination*. This coefficient is given as:

$$R^2 = r^2$$

R^2 has a range $0 \leq R^2 \leq 1$. The coefficient of determination is a kind of a percentual measure of the strength of the relation among the variables x and y. If R^2 is close to zero or low then the variation in x does to explain much of the variation in y. Vice versa if R^2 is high and close to one. R^2 is a very widely used measure, and we shall return to R^2 in the topic *regression* in the course Statistics II.

2. Example

Movies and Ratings
This example is taken from the exam February 2012, and it had a weight equal to 15 percent or 3 points.

When a new movie is released how large is the impact of the reviews in newspapers and other Medias on the amount of tickets sold? Does this guarantee that a lot of people go to the cinema and watch the movie?

This problem has been addressed by the Danish magazine on movies "Ekko" in the issue of December 2011. The magazine investigated the relation for movies specifically produced for teenagers.

The result of their investigation is shown in the table on the top of the next page. The table brings together the number of tickets sold y and the rating x taking values from 1 (worst) to 5 (best) calculated as the mean rating of the rations from four leading Danish newspapers. The investigation covers the 24 Danish produced movies specific for the teenage groups for the period 2002 to 2011.

In addition to the table the following information's were provided:

$$\bar{y} = 95.38 \quad \bar{x} = 3.48 \quad \sum_{i=1}^{24} y_i^2 = 514159.64 \quad \sum_{i=1}^{24} x_i^2 = 300.4 \quad \sum_{i=1}^{24} (x_i - \bar{x})(y_i - \bar{y}) = 676.22$$

A. Calculate the standard deviations for X and Y. **1P**
B. Calculate a measure for the degree of the strength among X and Y and provide an interpretation of the result. **2P**

Release	Title	Y: Tickets (1000)	X: Rating
2002	Slim, slam, slum	7.8	1.6
2003	Midsummer	121.3	3.4
	Football friends	78.2	3.6
	Hinterland	90.6	4.3
2004	Count to 100	23.1	2.4
	Terkel in disguise	375.8	4.4
2005	Off the road..	71.7	2.8
	Young Andersen	10.0	3.3
	Strings	15.7	3.8
2006	Super adult	106.5	3.6
2007	Fighter	52.2	3.5
	Rich Kids	107.0	2.6
2008	Two worlds	314.5	4.0
	You and me	101.8	3.6
	The journey to Saturn	401.0	4.0
2009	Love troubles	175.1	3.6
	See my dress	45.4	4.0
	Madly in love	27.4	4.0
2010	Hold me	46.0	3.2
	My favorite enemy	9.4	3.4
2011	Free fall	29.2	4.6
	Bora bora	75.4	2.8
	Skyscraper	1.1	3.6
	Magic in the air	3.0	3.3
Sum		2289.2	83.4

Solution

This exercise can be solved manually by use of the information provided above or by use of Excel. Initially consider the case with the manual calculation.

A)

In order to calculate the standard deviations insert the numbers such and obtain:

$$s_y = \sqrt{\frac{1}{24-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)} = \sqrt{\frac{1}{24-1}\left(514159.64 - \frac{(2289.2)^2}{24}\right)} = 113.41$$

and

$$s_x = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)} = \sqrt{\frac{1}{24-1}\left(300.4 - \frac{(83.4)^2}{24}\right)} = 0.68$$

Notice that I got the sum of the *x* and the *y* variable from the bottom line of the table.

B)
The coefficient of correlation is the best measure to show the degree of the strength of the relation among the two variables. Initially calculate the covariance.

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{676.22}{24-1} = 29.40$$

Alternatively the covariance can be calculated by use of the formula:

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{n-1} = \frac{8631.36 - 24 \times 95.38 \times 3.48}{24-1} = 29.40$$

Using the information found under question A, the correlation can be found as:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{29.40}{0.68 \times 113.41} = 0.3812$$

As expected a positive correlation can be observed.

So a good rating will increase the number of tickets sold. However this relation is not very strong.

Next let us move to the solution using **Excel**. Load in the information provided in the table above into Excel.

A)
The standard deviations can be used by the function *Descriptive Statistics.* However, remember that you have to explain the content of the formula.

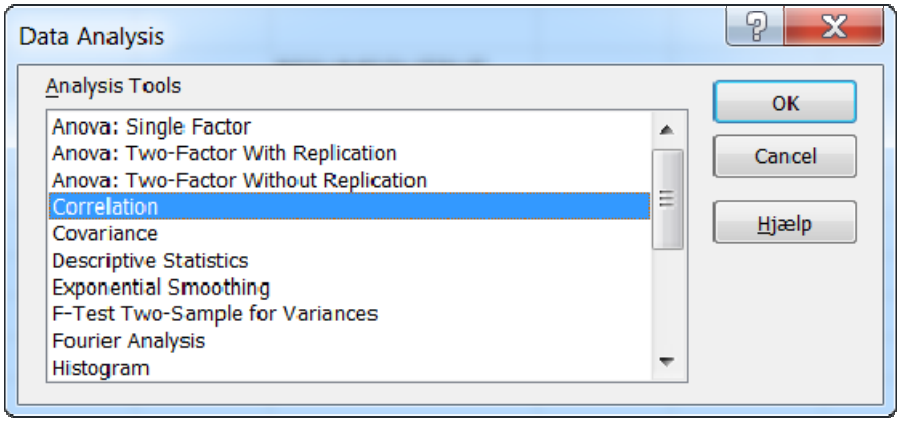
The full *descriptive analysis* looks as:

	Tickets	*Rating*
Mean	95.39	3.48
Standard Error	23.15	0.14
Median	61.95	3.6
Mode	#I/T	3.6
Standard Deviation	113.41	0.68
Sample Variance	12862.08	0.46
Kurtosis	2.61	1.32
Skewness	1.83	-0.87
Range	399.9	3
Minimum	1.1	1.6
Maximum	401	4.6
Sum	2289.24	83.4
Count	24	24
Confidence Level (95.0%)	47.89	0.29

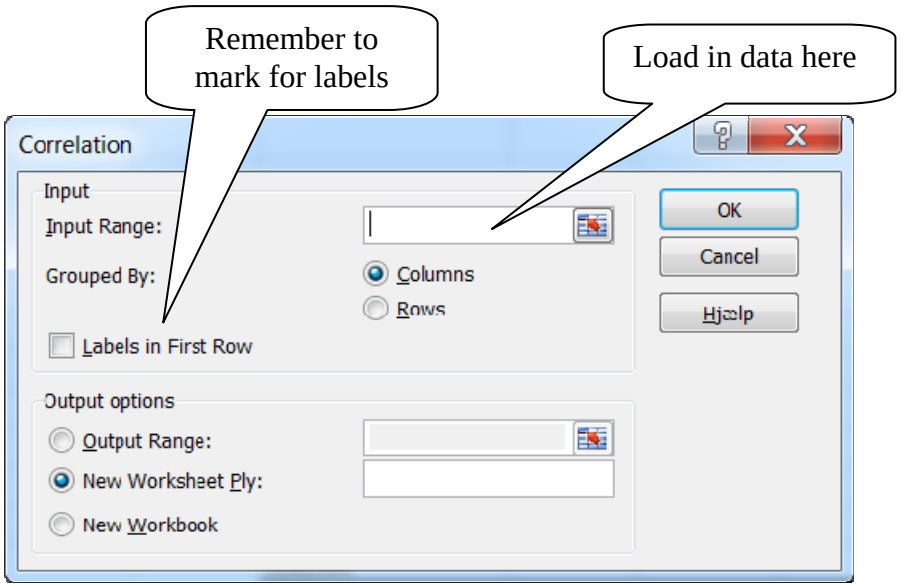
Remember that you need to explain the content in the table

B)

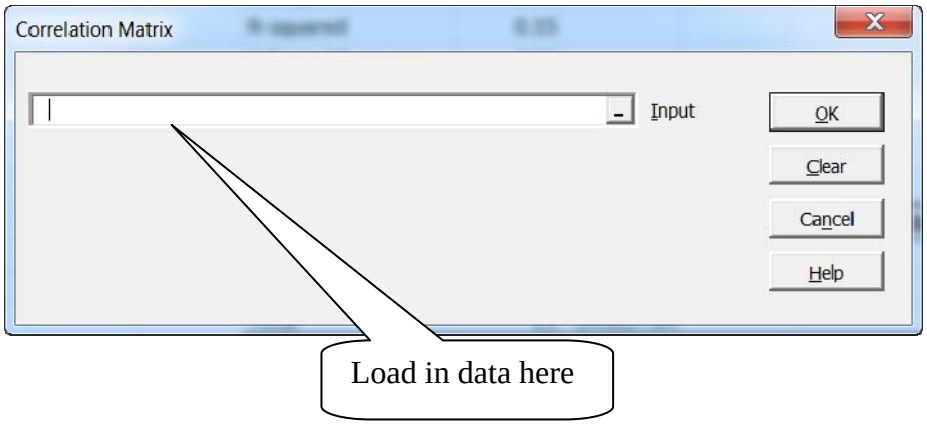
The coefficient of correlation can be found in the **Data Analysis** box as:



The dialog box for correlation now appear:



In the program **Megastat** use the frame *correlation/regression* and select *correlation matrix.* Then the following box appears. Mark in data, and obtain an answer similar to the answer found earlier.



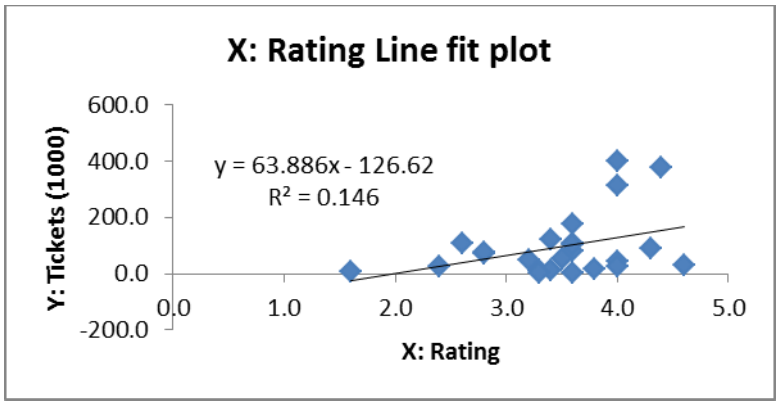
The output looks as:

Correlation Matrix

	Y: Tickets (1000)	*X: Rating*
Y: Tickets (1000)	1.000	
X: Rating	.382	1.000

24 sample size

Let us use Excel to set up a scatter plot of the data.



In this illustration I have set up data such that the rating is on the horizontal axis and the number of tickets sold is on the vertical axis.

I have calculated the line by the "add trend line" facility. Mark one of the blue points in the diagram and clicked on the left bottom of the mouse. Under "add trend line" I have marked "show equation in diagram" and "show R^2 in diagram.

The R^2 is equal to $R^2 = r^2 \Leftrightarrow R^2 = \sqrt{0.3812} \Leftrightarrow R^2 = 0.146$.

The equation explains 14.6 percent. So the rating explains 14.6 percent of the variation in the ticket sales. This is not very much, so the relation cannot be said to be strong. From the table it can be observed that many of the movies with high ranking not sells that many tickets. This could be due to marketing or the structure of the ownerships of the cinemas showing the movies.

Nils Karl Sørensen ©
Statistics I, BA-INT

Set 3: Probability and Distributions

by Nils Karl Sørensen

For the probability distributions, we need to use **Statistics Tables**. They can be found in BlackBoard under the folder *Statistics Tables.*

Outline

page

1. Probability and Statistics

The character Gladstone Gander from Donald Duck is amazingly lucky. He will not only win a lottery once, but *twice* within a period. This event is not likely, but it is possible! How possible? In order to assign how possible, we introduce the concept of probability.

The winner is chosen randomly from all the participants in the lottery. So drawing the winner of the lottery is the same as taking a sample of one lucky winner from the total population (all the participants).

Statistical inference is a method where we draw conclusions (inference) about a population based on a sample drawn from the population.

The population is also in probability theory referred to as the *space.* The space depends of the number of participants in the lottery. Relative to the case of Gladstone Gander it may be all citizens in Duckburg who participates in the lottery. They can buy a lottery coupon and participate in the lottery. All coupons are put into a box, and it is shaken, and one coupon is selected. So we use simple random selection to find the winner. If each participant only is allowed to buy one coupon and there are n coupons, then the probability to win is $1/n$. This is very similar to the concept of the frequency developed in the previous set of notes.

In general, define the probability as:

Probability = quantitative measure of uncertainty

A probability must be assigned to an outcome resulting from an *experiment.* Performing an experiment is something that results in an uncertain outcome. In the example above, the lottery is the experiment, and drawing the winner of the lottery is performing the experiment.

For a probability, two conditions should be met:

1. The probability assigned to each experimental outcome must be between 0 and 1. That is if *E* represents an experimental outcome and if *P(E)* represents the probability of this outcome, then $0 \le P(E) \le 1$.
2. The probability of all the experimental outcomes must sum to 1.

For example, if we toss a coin there are two experimental outcomes: Head (H) and tail (T). If the coin is fair, the probability of for example the event (E) of getting a head when the coin is tossed must be a half or 0.5. The probability associated with getting a tail is similar equal to 0.5. As 0.5 + 0.5 = 1, we can see that both conditions are met for our experiment.

How can we be sure that the coin is fair? We can replicate the experiment say 250 times. If the coin is fair, there should be around 125 heads and 125 tails. We can calculate the probability for heads for the experiment as:

$$P(H) = \frac{number\ of\ heads}{number\ of\ trials} \approx 0.5 \qquad \text{as the number of trials goes to infinity}$$

This statement is called *Cordanos rule*. The probability here is based on the observed outcome of the experiment. It is also called an *objective probability*. A probability can be accessed in two ways. Either as an *objective probability* or as a *subjective probability:*

1. *Objective probability:* This is based on calculated observations, symmetry of games of change or similar situations. This is also called the *classical probability*.

2. *Subjective probability:* This is on the other hand based on personal judgment, information, intuition, and other subjective information criteria. The area of subjective probability – which is relatively new, having first been developed in the 1930s – is somewhat controversial. Subjective probability is also called *personal probability*. Consider as an example the weather forecast. Here the analyst uses information from several weather forecast simulation models and then by personal judgment select a combination of the most likely to happen.

Whatever the kind of probability involved, the same set of mathematical rules holds for manipulating and analyzing probability.

To understand probability, some familiarity with sets and with operations involving sets is useful for our calculus.

Define:

A **set** is a collection of elements.

Elements can be anything that can be associated with a number. In the example it could be all participants in the lottery in Duckburg.

The **empty set** is the set containing *no elements*. This is called "**Ø**".
The **universal set** is the set containing *everything.* This is called **S**

Given the set **A**, we may define its **complement:**

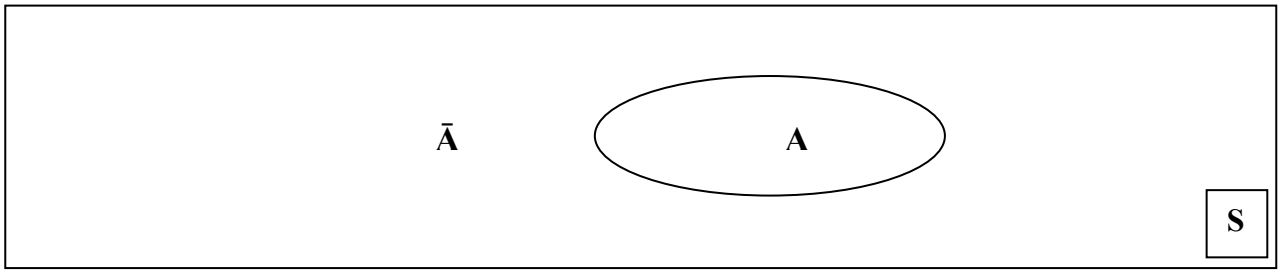
The **complement** of set **A** is the set containing all the elements in **S** not in **A**.

We can denote for any event *A* the probability *P(A)* satisfies:

$$0 \le P(A) \le 1$$

A ***Venn-diagram*** is a schematic drawing of sets that demonstrates the relationships between different sets.

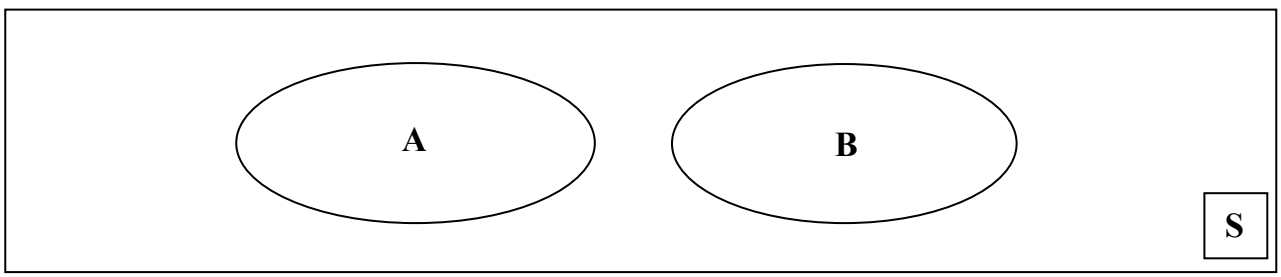
Let us define some situations. The first Venn-diagram illustrates the complement of an event.



Probability of a complement: $P(\overline{A}) = 1 - P(A)$

Relative to our example, **A** is the set of persons in Duckburg who participate in the lottery. The complement is all persons in Duckburg who not participate in the lottery. Finally, **S** is all citizens in Duckburg.

Now expand the situation. Introduce the lottery in the neighbor city to Duckburg called Calisota. In Calisota there is a lottery similar to the one in Duckburg. All the participants in each city can participate in their own lottery only. The persons who participate in the lottery is Calisota is denoted by **B**. The lotteries are exclusive. If **S** is all inhabitants in both cities then a Venn-diagram illustrating the mutually exclusive or disjoint sets (nothing in common) look as:



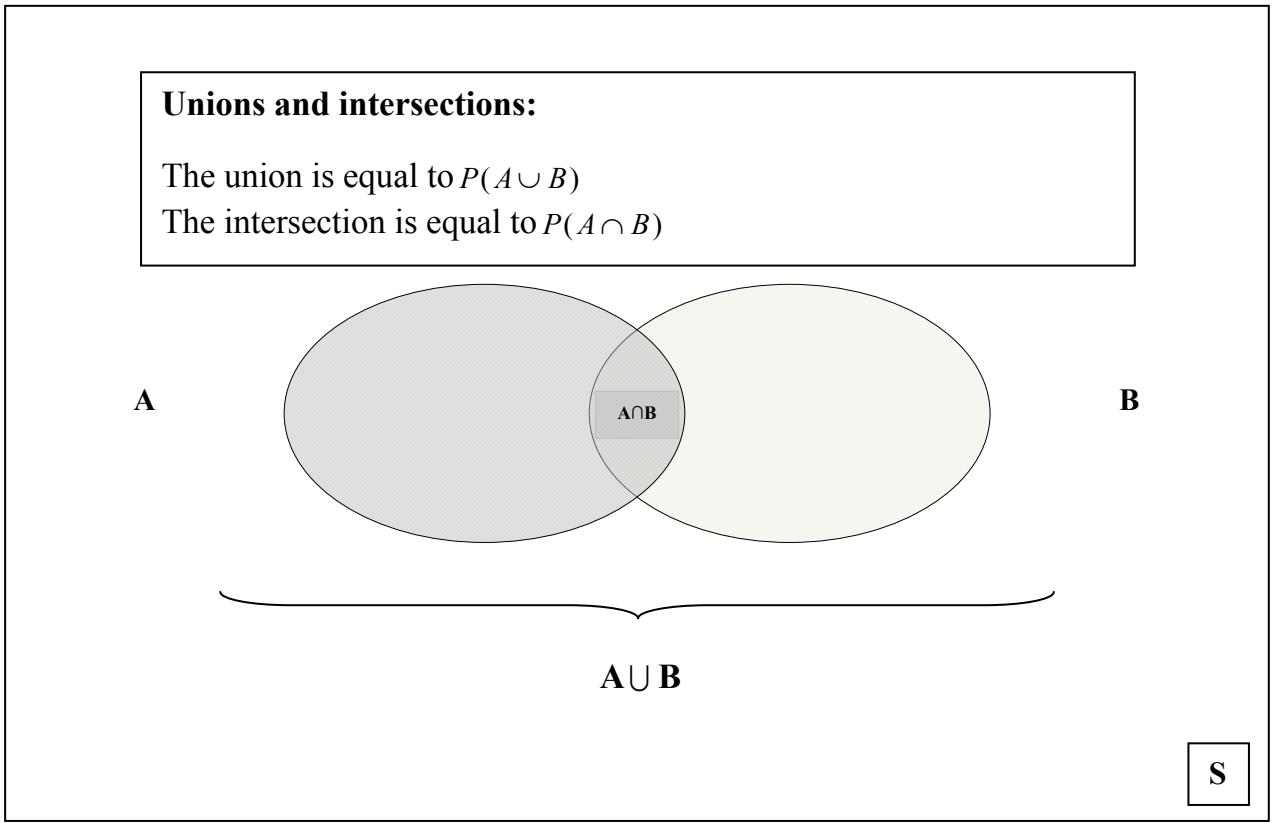
For mutually exclusive events $P(A \cap B) = 0$, and in this situation the probabilities can be added:

$$P(A \cup B) = P(A) + P(B)$$

The two Venn-diagrams above illustrate two extreme cases. Consider now what happened in the situations in-between. In such cases some relation will exist among **A** and **B.** In order to deal with such cases the concepts of *intersections* and *unions* has to be introduced:

Union	=	events that occurs in **A** or **B** (or both)
Intersection	=	events that occurs in **A** and **B**

The Venn-diagram illustrates the cases.



The union can be found as the summation of the outcomes of the two events:

$$P(A \cup B) = P(A) + P(B)$$

This is rule from the top of the page again! This is called the *addition rule*. What if this not is the case? For example it could be possible for the citizens in Duckburg and in Calisota to participate in the lotteries in both cities. In this situation the events are not mutually exclusive. As a result some citizens will participate in both lotteries. In this case the

intersection is $P(A \cap B) \neq 0$. The citizens in the intersection participate in both lotteries. The union must be restated as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

What happens? In order to find the union the intersection has to be deducted. Think of this in the following way: Take two pieces of A-4 paper – one in each hand, and call the paper to the right for A and the paper to the left for B. To find the union put them together, and let them have some kind of overlap. It is observed that the overlap is counted twice. In order to correct for this, one of the pieces of paper has to be cut such that there is no overlap. The piece of paper left out is equal to the deducted intersection.

The rule of addition may easily be complex. Consider for example a situation where a third city is introduced. Further citizens in all 3 cities are allowed to participate in the lotteries in all 3 cities. Denote the third city by **C.**

The union in this case is written as:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$
$$- P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Try for yourself to do this with 3 pieces of paper – quite complicated!

Above it was showed that the union of two sets could be found by summation if the intersection was equal to zero.

The *intersection* can be found as:

$$P(A \cap B) = P(A)P(B)$$

This is also called the *rule of multiplication*. It tells us that when **A** and **B** are *independent* we can obtain the probability of joint occurrence by multiplication of the probabilities of the events. This is an important rule.

Until now this section has been theoretical. Consider now an example.

Example on Education Programs

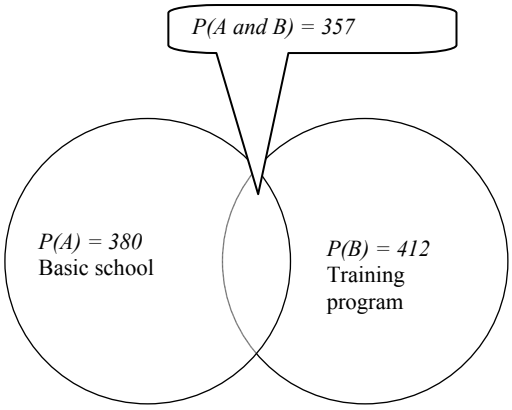
A firm has a total of 550 employees. Out of this 380 has basic school education whereas 412 has participated in a post school training program provided by the firm. In all 357 of the employees has a basic school education as well as the training program.

If an employee is randomly selected what is the probability that:

A. The employee has a basic school education?
B. The employee has participated in the training program?
C. The employee has both?

Solution
There are two types of education. Basic school denoted by **A** and training program denoted by **B.** Further there is the intersection of employees who has **A** as well as **B.**

The situation can be stretched out by a *Venn-diagram:*



The probabilities of having a given education can be stated as:

A.	*P(A):* Basic school	380 → P(A) = 380/550 = 0.691
B.	P*(B):* Training programme	412 → P(B) = 412/550 = 0.749
	Total	550

This may seem a bit confusing, but the point is that the two probabilities should sum to *more* than one, because they not are mutually exclusive.

There must then be an intersection of sets with basic school as well as the training programme.

$P(A \cap B)$: Persons with both educations equal: 357

We now use the law of union (addition rule) to find the union set.

C.
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{380}{550} + \frac{412}{550} - \frac{357}{550} = 0.691 + 0.749 - 0.649 = 0.791$$

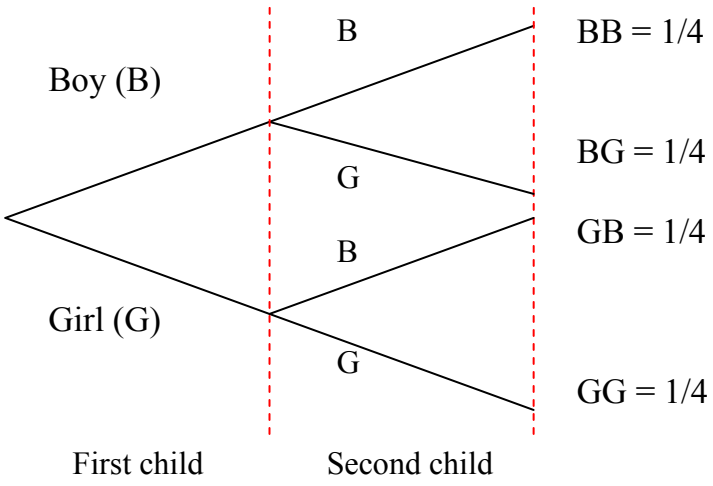
2. Tree Diagrams

In order to calculate probabilities by use of the classical probabilities, it is important to determine the *sample space.* This is an examination of all possible outcomes. In order to undertake such an investigation we use *tree diagrams.* Here it is attempted to set up a sequence of events *conditional* of each other. For example: First event A happens, then how likely is that event B will occur? Such a conditional probability is written as:

$P(A|B)$: The probability for A given that B already has occurred

In this way the *space* is determined for the conditional probability. Remember for above that if two events are independent then the probabilities can be multiplied.

Consider the example.

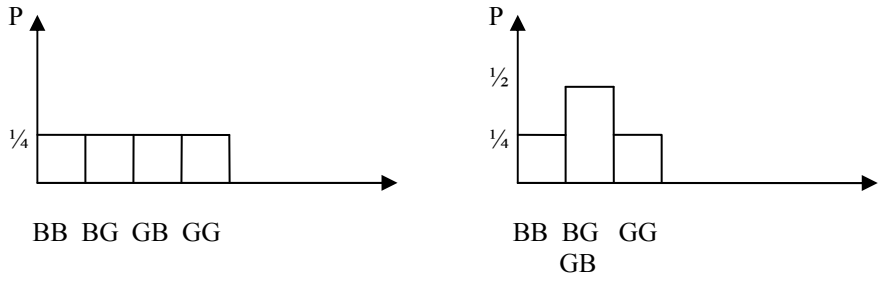


Here a couple plans to have two children. Approximately the probability of having a girl (G) or a boy (B) are equal, so *P(G) = P(B) = 0.5.* Let us assume that they have one at a time, so no twins!

The sample space is equal to: BB BG GB GG

With four outcomes are then: P(BB) + P(BG) + P(GB) + P(GG) = 1

Notice that the distribution of the outcomes is independent and uniform, see the left panel below. Let us now take a sample of 2 out of the 4 outcomes. Then the distribution will be as in the panel to the right. The diagram look a little bit like a normal distribution (bell-shaped)! We get back to this point later in this set of notes.



Example on Election Performance (Exam BA-INT February 2011, 10%)
An election to the parliament has taken place in a country. Out of the registered voters 70 percent participated and 20 percent of them gave vote for the party XYZ.

How large is the probability that a randomly selected voter has participated in the election, but has *not* given vote to the party XYZ.

Solution:
First write up the information stated in the exercise. Do this in order to get a clear picture on the problem.

- Election participation is equal to 70 percent of all registered votes. The probability to vote is P(V) = 0.7

The probability that is asked for has to be found *given* that a vote has been undertaken. Then is a *conditional* probability that has to be found.

- Given that a voter has participated in the election and voted for the party XYZ then the probability is P(XYZ | V) = 0.2

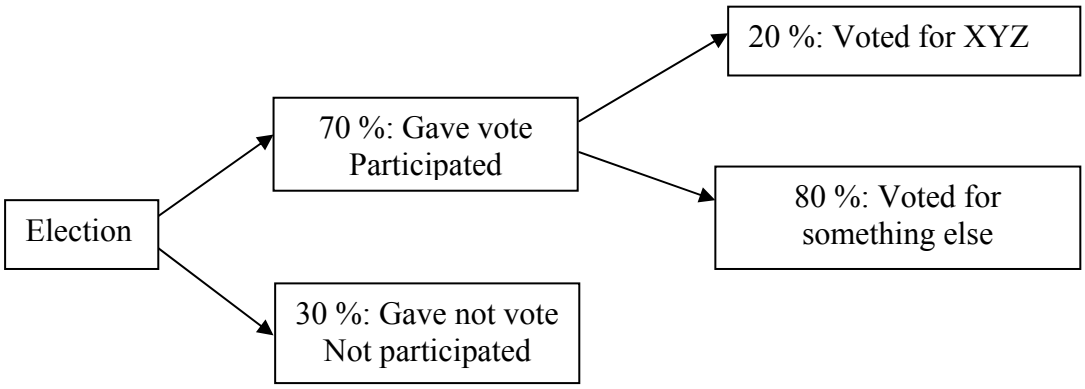
- Given that a voter has participated in the election and not voted for the party XYZ then probability is P(non-XYZ | V) = 1 – 0.2 = 0.8

The last probability is to have voted for another party. However, it has not been taken into account that a randomly selected person not has voted for any party, but is registered for the election.

$$P(\text{V and non-XYZ} \mid \text{V}) = P(\text{V}) \times P(\text{non-XYZ} \mid \text{V}) = 0.7 \times 0.8 = 0.56$$

The probabilities can be multiplied because election vote and election participation are independent event.

These considerations may seem messy! Consider as alternative to set up a tree diagram. See top next page.



So for participation, but not voted for XYZ then $0.8 \times 0.7 = 0.56$

3. Introduction to Probability Distributions

An event may have an uncertain outcome. For example:

- Who will win the next parliament or local election?
- How old will I be?
- Will the couple in the example above first get a boy or a girl?
- Will Gladstone Gander will the lottery in both Duckburg and Calisota?

All these events are *stochastic* and depend on chance. A probability can be accessed, i.e. a value ranging between zero and one to the outcome and the probability is a *random variable.*

- A **random variable** is an uncertain quantity whose value depends on chance.

- A **random variable** is a function of the sample space.

In the case with the children it is possible to write up the probabilities, and the outcomes are limited. This is not always the case. The statistical office can provide material on the expected lifetime by age, and that make it possible to provide an estimate of your lifetime. However, this is not certain, and for example an unexpected accident may occur tomorrow. There are many possibilities.

These considerations lead to a division of random variables into two classifications:

- A **discrete random variable** can assume at most a countable number of values
- A **continuous random variable** may take any value in an interval of numbers (i.e., it's possible values are uncountable infinite)

For the present, we shall introduce three probability distributions. The two discrete distributions: the Binominal and the Poisson distribution, and the continuous Normal distribution.

The Binominal and the Poisson distributions both take their point of departure in the most simple situation with two outcomes namely success or failure. The Binominal distributions consider a situation at a given point in time, whereas the Poisson distribution considers a process in time. The Normal distribution is a generalization of the Binominal distribution with many outcomes.

4. The Binomial Distribution

Let us go back, and consider the situation with 2 outcomes only namely success and failure. This could be:

- Will the couple in the example first get a girl or a boy?
- Which team will make the next goal in the football match?
- Will the car break down when you have invited the girlfriend out on a date?

How can we model these events, and what is the expected outcome of undertaking an experiment with two outcomes?

Bernoulli Random Variable
Very frequently a game has two outcomes or parameters namely success or failure. In such a case, we are working with a *Bernoulli random variable*. This situation is named in honor of the mathematician Jakob Bernoulli (1654−1705). The distribution of a Bernoulli experiment is given as:

x	$P(x)$
1 (success)	p
0 (failure)	$1-p$

We can calculate the expectation and the variance for this random variable by use of rules of expectation. Denote the expectation by E and the variance by V. The rules that can be derived from probability theory are:

$$\begin{array}{lll} E(X) & = 1\times p + 0\times(1-p) & = p \\ E(X^2) & = 1^2\times p + 0^2\times(1-p) & = p \\ V(X) & = E(X^2) - [E(X)]^2 = p - p^2 & = p(1-p) \end{array}$$

Often the quantity $(1-p)$, which is the probability of failure, is denoted by the symbol q so $V(X) = pq$.

Bernoulli Random Variable with Several Trials

In the real world, we often make several trials, not just one, to achieve one or more successes. Let us consider a situation with several Bernoulli-type trials. After all, the couple wanted two children and there will be many attempts on goal during the football match.

The coin that we worked with earlier is now tossed five times. Let H denote the number of heads that appear. If the coin is fair as before, we consider two equal outcomes. This is a *Bernoulli experiment.* The probabilities can in the present case be written as $p = (1-p) = q = ½$.

Now consider the problem:

What is the possibility that exactly 2 heads occur?

First, let us find the number of ways in which this event may occur. The sequences of five tosses of the coin that results in exactly 2 heads are:

TTTHH	TTHTH	THTTH	HTTTH
TTHHT	THTHT	HTTHT	
THHTT	HTHTT		
HHTTT			

There are in all 10 such sequences. Now we need to find the probability of each particular sequence. Owing to the independence of the five trials, the probability of a particular sequence, say, the sequence HHTTT, is the sum of the five events *ppqqq*. Since the order of appearance of heads or tails does not affect the probabilities, we see that this probability p^2q^3 is the probability of any one of the 10 sequences. Since there are 10 possible sequences leading to the value $H=2$, we multiply p^2q^3 by 10, giving us $P(H=2) = 10p^2q^3$. With the probabilities of a fair coin we obtain $P(H=2) = 10(½)^5 = 0.3125$.

It is possible to compute the probabilities of the other event in a similar way. We note that:

- The probability of any *given sequence* of x successes of n trials with probability of success p and failure q is equal to

$$p^x q^{n-x}$$

- The number of different sequences of n trials that result in exactly x successes is equal to the number of choices of x elements out of a total of n elements. This is equal to the number of combinations:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

On a pocket calculator combinations are labeled **nCr**. On a Texas **TI-84 pocket calculator** this is found by tasting ”MATH”. Then ”PRB” and so 3:nCr. The expression ”!” is faculty. This is the multiplicative sum of a given number. For example 4! = 1×2×3×4 = 24.

Consider an example. A group of 10 students has to select a team to arrange the Christmas event. In how many ways can this be done?

Use the formula:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{3628800}{6 \times 5040} = 120$$

There are then 120 combinations do undertake the formation of the team for the Christmas event. On the **TI-84 pocket calculator** the number can also be found directly. To find 10! First tast 10, and then ”MATH” → ”PRB” → 4:! → ENTER → 3628800.

The number of combination is found by first tasting 10, and then ”MATH” → ”PRB” → 3: nCr → 3 → ENTER → 120.

The Binominal Distribution
We can state the Binomial distributions as:

$$P(x) = \binom{n}{x} p^x q^{n-x} \qquad \text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

the term $\binom{n}{x}$ denotes the number of combinations (nCr), and “!” is the factorial. p is the number of success in a single trial, $q=(1-p)$, n is the number of trials, and x is the number of successes. The Binomial distribution is tabulated in the **Statistics Tables** in **Blackboard** Mean and variance is given as under the Bernoulli experiment, but now multiplies with the n trials. So:

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = npq$$
$$\sigma = SD(X) = \sqrt{npq}$$

The Bernoulli or Binominal distribution has the following characteristics:

- The experiment consists of *n* identical trials
- Each trial has two outcomes *p* (success) and *q* (failure)
- The probabilities are constant from trial to trial
- The trials are independent. This mean that we are drawing a little sample from a very large population

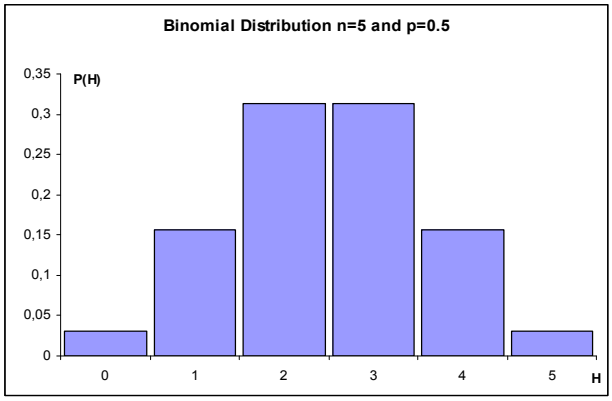
Let us calculate the probabilities in our example with *n=5*:

P(x)	Combinations (nCr)	Probabilities	P(H)	Cumulative P(H)
0	5!/[0!×(5−0)!] = 1	$(½)^0(½)^5 = 0.031$	0.031	0.031
1	5!/[1!×(5−1)!] = 5	$(½)^1(½)^4 = 0.031$	0.156	0.187
2	5!/[2!×(5−2)!] = 10	$(½)^2(½)^3 = 0.031$	0.313	0.500
3	5!/[3!×(5−3)!] = 10	$(½)^3(½)^2 = 0.031$	0.313	0.813
4	5!/[4!×(5−4)!] = 5	$(½)^4(½)^1 = 0.031$	0.156	0.969
5	5!/[5!×(5−5)!] = 1	$(½)^5(½)^0 = 0.031$	0.031	1.000

This is a symmetric distribution with mean, variance and standard deviation:

$$\mu = E(X) = np = 5 \times ½ = 2.5$$
$$\sigma^2 = npq = 5(0.5)^2 = 1.25$$
$$\sigma = \sqrt{npq} = \sqrt{5(0.5)^2} = 1.12$$

If *p=0.5* the distribution is symmetric. If *p>0.5* the distribution is skewed to the left. If *p<0.5* the distribution is skewed to the right. Let us now graph the distribution:



The probabilitity can also be calculated by use of a **TI-84 pocket calculator.** This task is undertaken as follows. Use the tast "2nd" and then "DISTR". Under A: Binompdf*(n,p,x)* → "ENTER". Insert now the numbers. This is: A: Binompdf*(5,0.5,2) = 0.3125.* As expected.

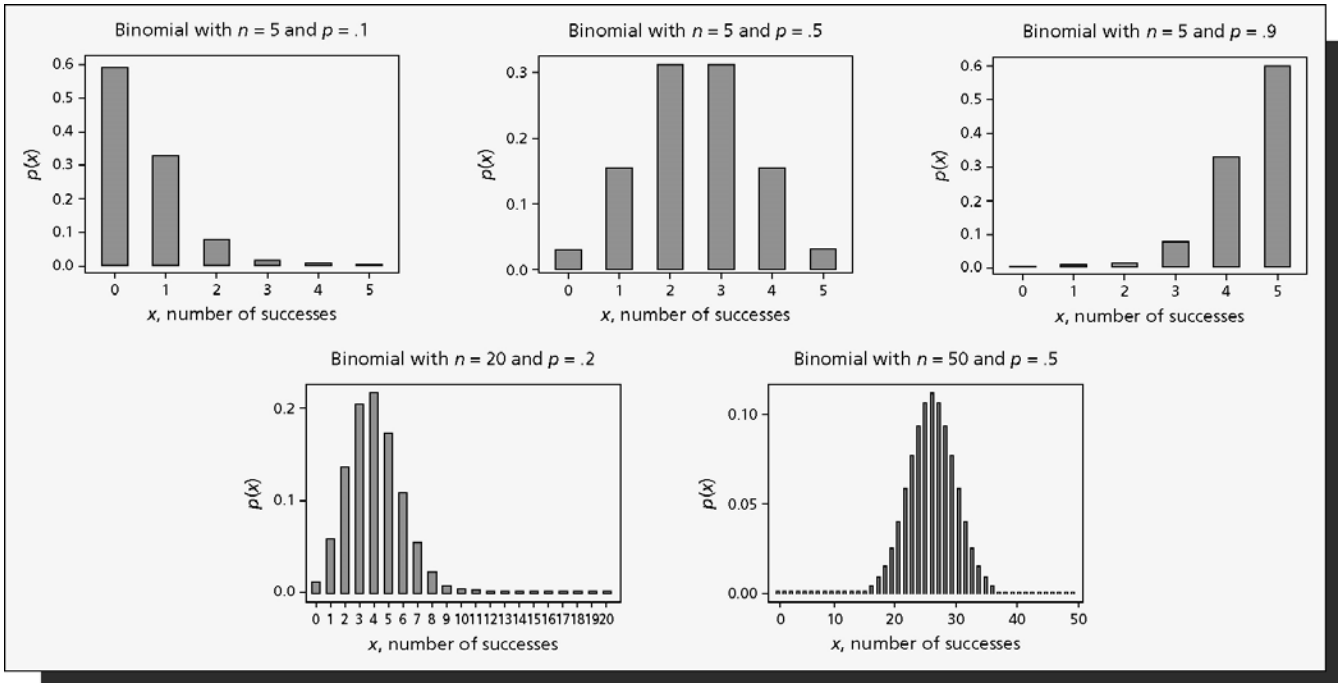
Use alternatively the **Statistics Tables.** The tabulation of the Binominal distribution is the first in the set of tables. Look at page 2 mid in **Statistics Tables**, and find the table on the next page.

n = 5					*p*						
X ↓	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	
0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313	5
1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563	4
2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125	3
3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125	2
4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563	1
5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313	0
	0,95	**0,90**	**0,85**	**0,80**	**0,75**	**0,70**	**0,65**	**0,60**	**0,55**	**0,50**	**X ↑**

Observe the way that the table is build. First we find the number of trials. This is **n=5.** Next, we find the probability of a "head". This is equal to **p = 0.5.** Finally, we find for exactly two heads, so **X = 2.** Here the probability is equal to **0.3125**. The diagram above is just the probabilities from this column.

What if p>0.50? Then the table should be read reversed. That is from right bottom to left. For example if p=0.75 and X=3. Look at the bottom line and find 0.75. Move up in the table and at X=3 find the probability 0.2637.

Finally, let us consider the shape of some Binominal distributions:



Notice that:

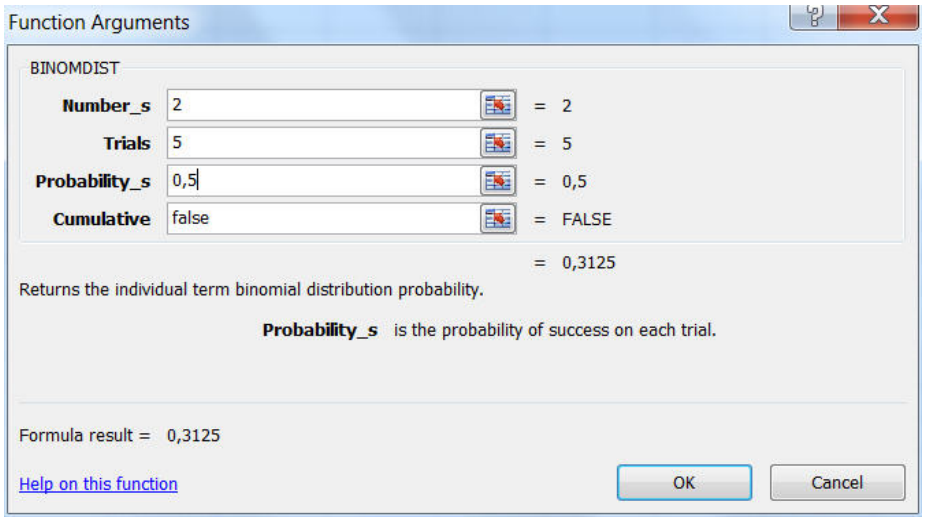
- The Binominal distribution becomes more symmetric as *n increases* and as $p \to 0.5$. (Lower right panel)
- The Binominal distribution becomes *skewed to the right* as $p < 0.5$.

(Upper left panel)

- The Binominal distribution becomes *skewed to the left* as $p > 0.5$.
  (Upper right panel)

As *n* increases, the binominal distribution will approximate the normal distribution. We will return to this below. The Binomial distribution can also be found in Excel by typing:

Formulas / Insert function / statistical /biomdist



Here, I have found the probability in the example above.

It is also possible to use Excel to find the cumulative probability. Consider the options:

- An example of the discrete probability is *P(X=4)* ("FALSE" in Excel)
- An example of the cumulative probability is *P(X≤4)* ("TRUE" in Excel)

On the pocket calculator to get the cumulative probability B: Binomcdf*(n,p,x)*.

5. The Normal Distribution

The normal distribution is an important continuous distribution because both the Binominal and the Poisson distribution can be approximated to it. Further, it has turned out to be a very stable functional form.

The normal distribution is also called the *Gauss Distribution* after the German mathematician Carl Friedrich Gauss (1777−1855). Gauss showed that the mathematical functional form given below was the most optimal to approximate random errors obtained by the least squared errors method. We shall return to this method in Statistics II in the notes on *Regression Theory*. In 1802, Gauss[1] used this method to very precisely estimate the orbit of the just newly discovered asteroid *Ceres´*. Further, the normal distribution is extremely useful when working with the Central Limit Theorem. We shall return to this issue in the next set of notes.

Gauss has been a very central scientist in the development of mathematics and statistics. Today he would have been a candidate to the Nobel Prize, and he appeared on the old German D-mark notes as well as on stamps in the DDR!

For a normal distribution with mean μ and standard deviation σ, the probability density function *f(x)* is given by the complicated formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad \text{for } -\infty < x < +\infty$$

Where *e* is the natural base logarithm. Illustration is given on the next page. The distribution has mean equal to zero and standard deviation (as well as variance) equal to one. This can be written as $Z \approx N(0,1^2)$. For an variable *X* this is frequently written as

[1] Although Gauss was a brilliant mathematician he was also a person of a very cumbersome nature. Living also in Germany at the same time Alexander von Humboldt was also a famous and incredible scientist who travelled over most of the planet. German writer Daniel Kehlmann has written a very interesting and funny novel entitled "Die Vermessung der Welt" where he puts the two very different personalities together in a period under political change. The book has been translated into most European languages and has sold in several million copies. In 2012 the book was tuned into a movie directed by Detlev Buck.

$$X \approx N(\mu_x, \sigma_x) \qquad \text{or} \qquad X \approx N(\mu_x, \sigma_x^2)$$

The rules from the notes on expected values of discrete random variables are also valid for the normal distribution.

Any variable X can be transformed to Z by use of the important formula

$$Z = \left(\frac{X - \mu}{\sigma} \right)$$

The transformation takes us from a random variable X with mean μ and standard deviation σ to the standard normal random variable Z.

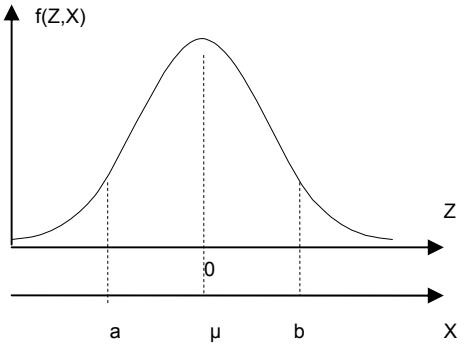
Based on the transformation probabilities can be found by use of **Statistics Tables** page 9 as the cumulative areas under the standard normal curve. We can also find the values by use of Excel or be a pocket calculator. See later!

Transformation formulas of X to Z, where a and b are numbers are given by the formulas

$$P(X < a) = P\left(Z < \frac{a - \mu}{\sigma} \right)$$

$$P(X > b) = P\left(Z > \frac{b - \mu}{\sigma} \right)$$

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma} \right)$$



The transformation from X to Z can also be undertaken *inversed* This takes us from the standard normal random variable Z to the random variable X with mean μ and standard deviation σ. The inverse transformation is given by the formula

$$X = \mu + Z\sigma$$

The transformation formulas enables us for given numbers a and b to find the associated probabilities under the normal distribution curves.

Example: The Price of Gasoline

During the last couple of years, the market for gasoline has been characterized partly by rising oil prices and partly by price warfare on the product. In other words, driving a long way to get the good offers has often been worthwhile!

During the spring of 2006 a price-conscious car owner, who always seeks to fill up his car with gasoline at the lowest possible price, gathered information on the price per litre unleaded 95 octane's during 25 visits to the gas station. The mean price was 9.95 DKK and the standard deviation was 0.30 DKK. Now assume that the price is normally distributed with this mean and this standard deviation.

On this basis, answer the following questions:

A. Once the price of gasoline was 10.40 DKK per litre. What is the probability that the price will be higher at the car owner's next visit to the gas station?
B. Once the car owner was lucky to buy gasoline at 9.65 DKK per litre. What is the probability that the price will be lower at his next visit to the gas station?
C. What is the probability that the car owner will fill up at a price of between 9.80 DKK and 10.25 DKK at his next visit to the gas station?
D. Suppose now that the car owner expects a lower price 25 % of the times he arrives to the gas station. What is the minimum price he will accept given the mean and the standard deviation?

Solution to Example

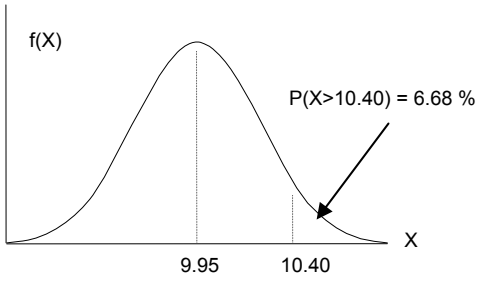
(The statistical material is my own observations collected on various trips from Haderslev where I live to the SDU campuses of Esbjerg or Sønderborg where I work)! We consider a normal distributed dataset with *N(9.95;0.30)*. The size of the sample is *n=25* (at this moment, we do not need this information, but we shall use it later).

A)

Once the price was 10.40 DKK per liter. What is the probability that the price will be higher at the car owner's next visit to the gas station?

$$P(X > 10.40) = P(\frac{X-\mu}{\sigma} > \frac{10.40 - 9.95}{0.30}) = P(Z > 1.5) = 1 - 0.9332 = 0.0668\ (= 6.68\ \%)$$

Illustration:



How did I got the probability? I used the **Statistics Tables** on page 9. I need to find the probability associated to Z = 1.5. Below find an extract of the table:

Z	0,00	0,01	0,02	0,03	0,04	…	0,09
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	…	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	…	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	…	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	…	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	…	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	…	0,9706

At Z = 1.5 we find the probability 0.9332. Consider another example: What if Z = 1.73? Then we first find the row with Z = 1.7. Next find the second digit under 0.03 in the head of the table. Finally find the probability at Z = 1.73 as 0.9582 as indicated by the arrows in the table.

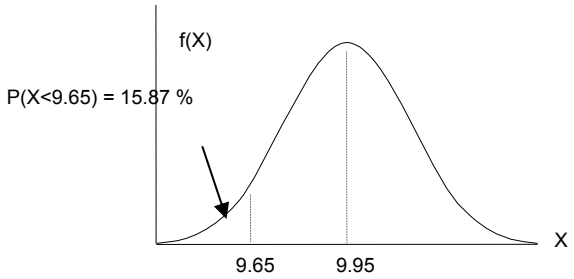
As an alternative one could find the probability by use of a **TI-84 pocket calculator.** Use tast "2nd" and then "DISTR". Under 2: Normalcdf*(low,high,μ,σ)* → "ENTER". Here is *low* lower bound, and *high* higher bound.

In the example the sequence will look as: 2: Normalcdf*(0,10.40,9.95,0.30)* = 1−0.9302 = 0.068. The probability is calculated up to the point 10.40 DKK. This probability is equal to 0.9302. This probability is subtracted from 1, and the answer to the exercise is obtained.

B)
Once the car owner was lucky to buy gasoline at 9.65 DKK per liter. What is the probability that the price will be lower at the next visit to the gas station?

$$P(X < 9.65) = P(\frac{X-\mu}{\sigma} < \frac{9.65-9.95}{0.30}) = P(Z < -1) = 0.1587\ (=15.87\ \%)$$

Illustration:



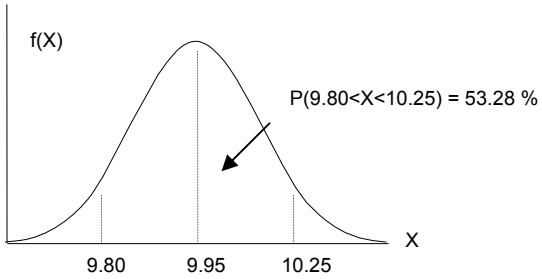
Again we find the value of Z by use of the table as above.

The table on page 9 in **Statistics Tables** is used as described above. On the **TI-84 pocket calculator** the following sequence is undertaken: "2nd" and then "DISTR" → 2: Normalcdf*(0,9.65,9.95,0.30)* = 0.1587.

C)
What is the probability that the car owner will fill up at a price of between 9.80 DKK and 10.25 DKK at his next visit to the gas station?

$$P(9.80 < X < 10.25) = P(\frac{9.80 - 9.95}{0.30} < \frac{X - \mu}{\sigma} < \frac{10.25 - 9.95}{0.30}) = P(-0.5 < Z < 1) = 0.8413 - 0.3085 = 0.5328\,(= 53.28\,\%)$$

Illustration:



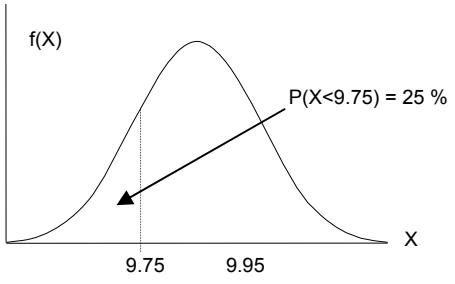
The table on page 9 in **Statistics Tables** is used as described above. On the **TI-84 pocket calculator** the following sequence is undertaken: "2nd" and then "DISTR" → 2: Normalcdf*(9.80,10.25,9.95,0.30)* = 0.5328.

D)
We solve the function inverse and use the equation given earlier. Consider the probability

$$P(X < a) = P(\frac{a - 9.95}{0.30}) = 0.25 \Leftrightarrow a = 9.95 + (-0.67)0.30 \Leftrightarrow a = 9.75$$

Here I found the probability 0.25 "inside" the table of the normal distribution, and then found the corresponding value of Z *(Z = −0.67)*. Finally, I solved for *a*.

A graphic illustration of question D can be found below:

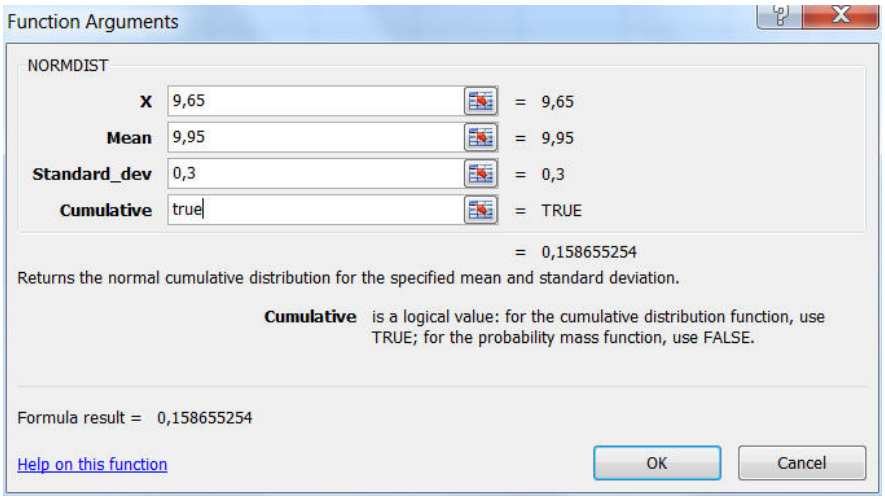


Here the **Statistics Table** page 9 is used "reversed". I have tried to illustrate this by use of the table below. We go "inside" the table, and find the probability associated with $Z = -0.67$. First find the row equal $Z = -0.6$. Then find the column equal to *0.07* (the second digit). At the intersection the probability is found to equal 0.2514.

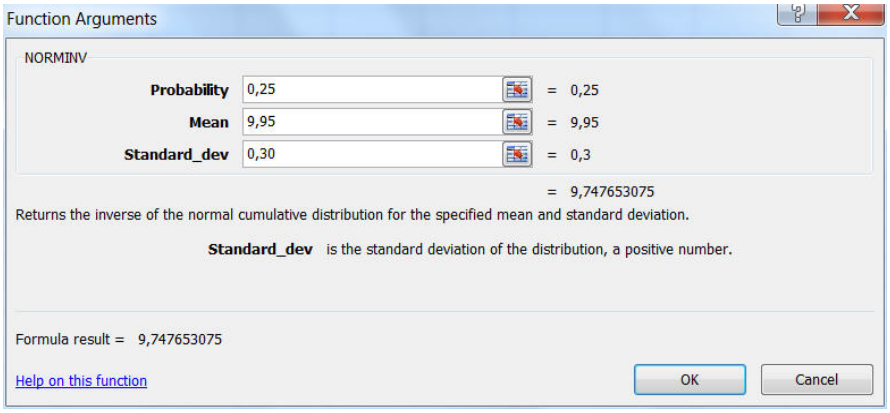
Z	0,00	…	0,05	0,06	0,07	0,08	0,09
-0,7	0,2420	…	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	…	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	…	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	…	0,3264	0,3228	0,3192	0,3156	0,3121

As above it is possible to find the probability by use of the **TI-84 pocket calculator.** Use "2nd" and then "DISTR". Under 3: InvNorm*(area,μ,σ)*. Where *area* denotes the probability. In the example use: 3: InvNorm*(0.25,9.95,0.30)* = 9.7476 ≈ 9.75. As expected!

We can also find the probabilities by use of Excel. Type **Formulas / Insert Function / Statistical / Normdist** and obtain the following:



Finally, we can use **Formulas / Insert Function / Statistical / Norminv** to find the answer to question D. Here is the screenshot:



6. Normal Approximation of Binominal Distribution

By use of the normal distribution, we can approximate binominal distributed statistics. Here the two distributions are considered by turn, but they can actually be related to each other.

This is most easily undertaken by substitution of the formulas for mean and variance for the binominal formula into the formula just given for finding the probability between to values *a* and *b*. For the binominal formula we have that:

$$E(X) = \mu = np \qquad \text{and} \qquad SD(X) = \sqrt{npq}$$

Then obtain

$$P(a < X < b) = P\left(\frac{a - np}{\sqrt{npq}} < Z < \frac{b - np}{\sqrt{npq}}\right)$$

This will hold for a binominal distribution with $n > 50$ and p ranging between $0.1 \le p \le 0.9$.

For smaller samples use instead for moderately sized samples n between 20 and 50:

$$P(a < X < b) = P\left(\frac{a - np - 0.5}{\sqrt{npq}} < Z < \frac{b - np + 0.5}{\sqrt{npq}}\right)$$

The implication of this formula is that the "tails" will be smaller. Finally, if p is very small or large use the Poisson approximation instead.

7. The Poisson Distribution

For example in process control, a time horizon is frequently considered for a discrete random variable. In such a case, we can consider a Poisson distribution developed by the French mathematician Simon Denis Poisson (1781−1840).

Assumptions:

- The probability that a given event occurs is proportional in time
- If the time interval is small the associated probability is small.
- The probability for an event to occur is independent of the initial time period considered.
- The probability for an event to occur in a given time period is independent of how many times it has occurred in the past.

The Poisson distribution can be stated as:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad \text{for } x=0,1,2,\ldots,n$$

Here μ is the mean (or expected) number of occurrences of the event in the specified interval, and e = 2.71828…. is the base for natural logarithms.

The use of the natural logarithm relates to the involvement of a time preference. We divide by *x!* in order to normalize.

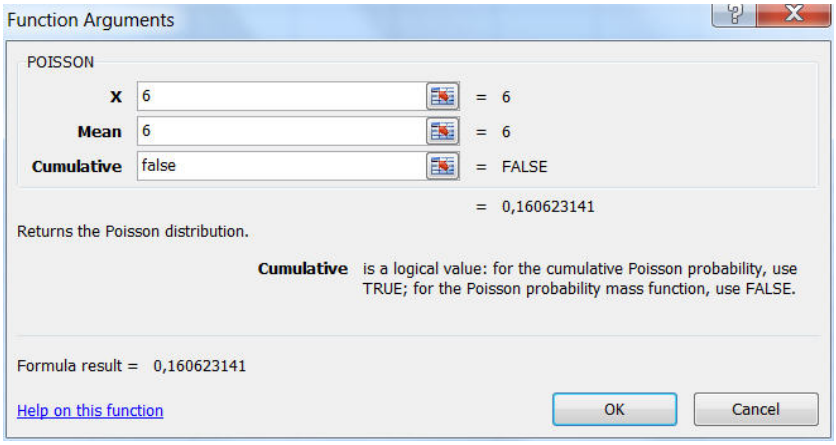
If $n>20$ and $p \leq 0.05$ the Binominal Distribution is a good approximation for the Poisson distribution. Notice that we can use the mean from the Binominal Distribution $\mu = np$ and substitute into the expression above. See also Section 8 of these notes.

Tables of probability distributions can be also found in **Statistics Tables** pages 6 – 8. We can also use Excel as usual by typing: **Formulas/Insert function/statistical/Poisson**. We need to type the value of the stochastic variable *X*, the mean μ and to judge on if we want cumulative distribution or not. A screenshot is found below.

Mean, variance and standard deviation is given by:

$$\mu = E(X) = \mu \qquad \sigma^2 = \mu \qquad \sigma = \sqrt{\mu}$$

In Excel the box for the Poisson distribution is shown on the next page. In the case shown, I have assumed a mean equal to 6, and I assumed that the number of occurrences also is 6, such that we find the probability of *P(X=6)*.

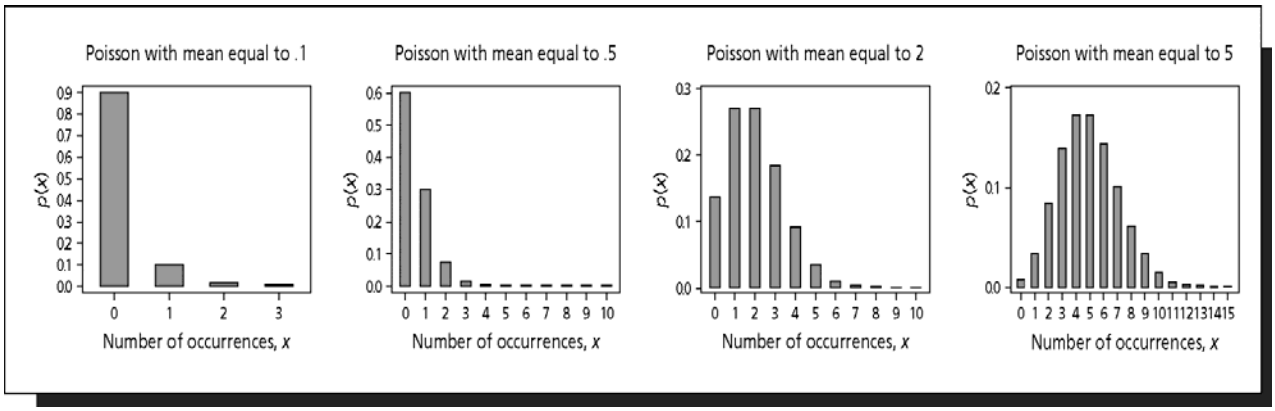


The Poisson distribution can also be found on the **TI-84 pocket calculator.** Use "2nd" and then "DISTR". Under C: Poissonpdf*(μ,X)* → "ENTER". In the example above *μ=6* and *X=6,* so C: Poissonpdf*(6,6) = 0.1606.* As expected! If we want the cumulative probability use D: Poissonpdf*(μ,X)* instead.

Look at the tabulation in the **Statistics Tables** page 7 mid extract:

			μ	
X	**5.5**	**6.0**	**6.5**	**7.0**
0	0.0041	0.0025	0.0015	…
1	0.0225	0.0149	0.0098	…
2	0.0618	0.0446	0.0318	…
3	0.1133	0.0892	0.0688	…
4	0.1558	0.1339	0.1118	…
5	0.1714	0.1606	0.1454	…
6	0.1571	0.1606	0.1575	…
7	0.1234	0.1377	0.1462	…
…	…	…	…	…
…	…	…	…	…
23	0.0000	0.0000	0.0000	0.0000

Finally, consider the distributions below. It is observed that the distributions become more "normal" as *n* and *μ* increases



Example: Waiting Time at the Drive-in at a Bakery

During the past decade, baker's with a "drive-in" facility has become increasingly popular in the region of South Jutland staring in the city of Haderslev. A "drive-in" facility for example makes it easy for families with children to buy bread, cakes etc.

(This is my local bakery! The newspaper "Politiken" had a very funny article dated November 22nd 2008 on the situation of being a "ghost driver" in the new drive-in").

Daily observations at the "drive-in" undertaken by the staff of the baker has revealed that over an independent interval equal to 10 minutes, a number equal to 6 drivers will arrive at the "drive-in". Assume that this process follow a Poisson distribution.

A. Denote the mean and variance for the distribution, and set up a histogram for the distribution

B. What is the probability that

- Exactly 6 drivers arrive during a period of 10 minutes?
- 4 or less drivers arrive during an interval of 10 minutes?
- 12 or more drivers arrive during an interval of 10 minutes?

Solution

A)

The big problem with exercises on the Poisson distribution is to find the ***time interval***. For the present problem we set $\mu = 6$ and define the time interval as 10 minutes.

The mean, variance and standard deviation of the distribution is given as:

$$\sigma^2 = \mu = 6 \qquad \Rightarrow SD = \sqrt{\sigma^2} = \sqrt{6} = 2.45$$

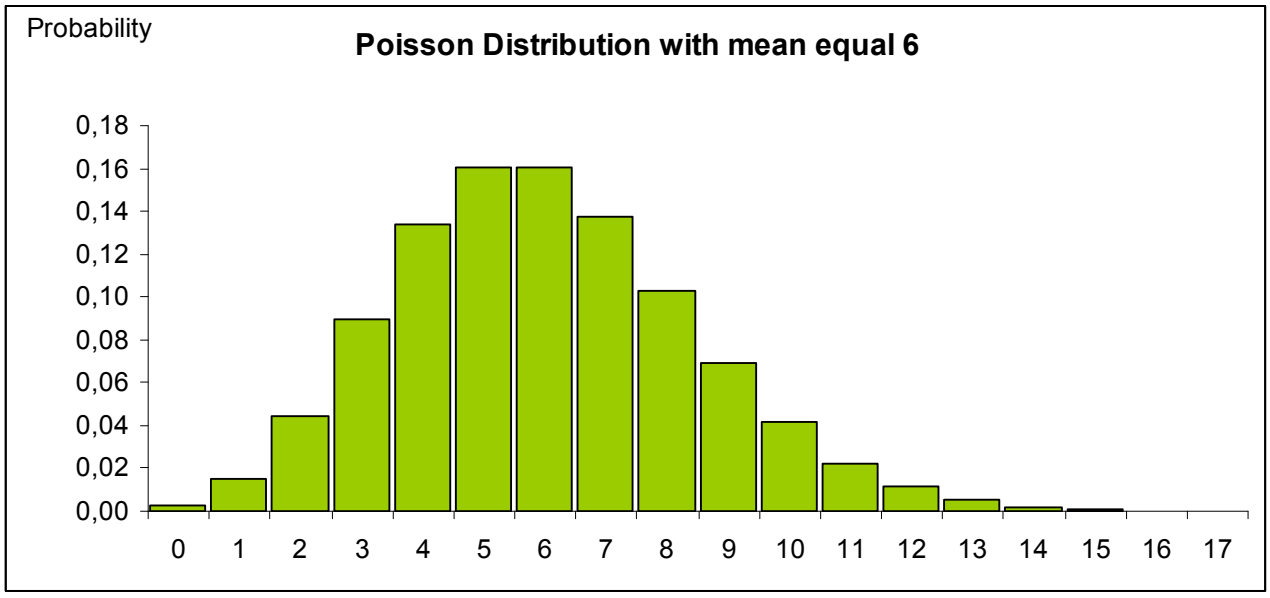
The mean and variance are by definition equal. This is so because we are working with a process in time.

Let us for illustrative purposes provide a little table of the probabilities and also set up a histogram. The probabilities are found in the **Statistics Tables** page 7 mid (second column of numbers with probability numbers).

x	0	1	2	3	4	5	6	7	8
P(x)	.0025	.0149	.0446	.0892	.1339	.1606	.1606	.1377	.1033

x	9	10	11	12	13	14	15	16	17
P(x)	.0688	.0413	.0225	.0113	.0052	.0022	.0009	.0003	.0001

And the histogram. We obtain (a close to symmetric) distribution:



B)
The three probabilities can be found either by use of table of the distribution (see above or en **Statistics Tables)** or by calculus and application of the formula just given at the beginning of this section.

- Exactly 6 drivers arriving during an interval of 10 minutes:

$P(x=6)$: $P(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{6^6 e^{-6}}{6!} = \frac{115.65}{720} = 0.1606$ (as expected from above)!

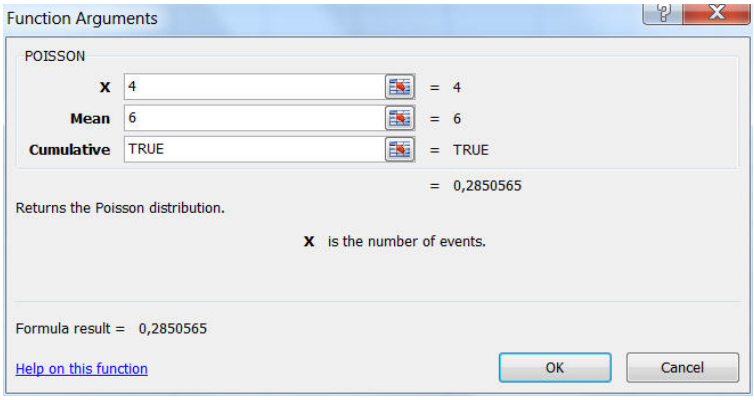
This result is also shown in the screenshot earlier!

- Less than 4 drivers arriving during an interval of 10 minutes:

$P(x \leq 4) = P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4)$

$$P(x \le 4) = \frac{6^0 e^{-6}}{0!} + \frac{6^1 e^{-6}}{1!} + \frac{6^2 e^{-6}}{2!} + \frac{6^3 e^{-6}}{3!} + \frac{6^4 e^{-6}}{4!} = .0025 + .0149 + .0446 + .0892 + .1339 = 0.2851$$

By use of Excel I obtain the screenshot:



Notice, that I in this case need the cumulative functional form "TRUE" in the box.

On the **TI-84 pocket calculator** for the cumulative probability use Use "2nd" and then "DISTR". Under D: Poissoncdf*(6,4)* → "ENTER" and obtain 0.2851. As expected!

- 12 or more drivers arrive during an interval of 10 minutes:

$$P(x \ge 12) = P(x=12) + P(x=13) + P(x=14) + P(x=15) + P(x=16) + P(x=17)$$
$$= 0.0113 + 0.0052 + 0.0022 + 0.0009 + 0.0003 + 0.0001 = 0.02$$

On the **TI-84 pocket calculator** for the cumulative probability use Use "2nd" and then "DISTR". Under D: Poissoncdf*(6,11)* → "ENTER" and obtain $1 - 0.9799 = 0.021$. As expected!

PS: Notice that I only go to 11. That is because we have to find $P(x \ge 12)$, so the probability for 12 is included in the number that have to found in the residual.

8. Poisson Approximation to the Binominal Distribution

The computation of probabilities for a binomial distribution can be tedious if the number of trials *n* is large. A mentioned above an approximation than be used in such a case. The probability of function of the approximating distribution can be obtained by substitution of the mean from the binominal distribution $\mu = np$ into the probability of the Poisson distribution:

$$P(x) = \frac{(np)^x e^{-np}}{x!} \qquad \text{for } x=0,1,2,\ldots,n$$

Example
An analyst predicted that 3.5 % of all small corporations would file for bankruptcy in the coming year. For a random sample of 100 small corporations, estimate the probability that at least 3 will file for bankruptcy in the next year, assuming that the prediction is correct.

Solution
The distribution of the number *X* of fillings for bankruptcy is binomial with $n=100$ and $p=0.035$. The mean is $\mu=np = 100\times(0.035) = 3.5$. We need to find $P(X \geq 3) = 1 - P(X \leq 2)$. We use the Poisson distribution, and find:

$$P(0) = \frac{3.5^0 e^{-3.5}}{0!} = 0.0302\,;\; P(1) = \frac{3.5^1 e^{-3.5}}{1!} = 0.1057\;;\; P(2) = \frac{3.5^2 e^{-3.5}}{2!} = 0.1850$$

Thus, $P(X \leq 2) = P(0) + P(1) + P(2) = 0.0302 + 0.1057 + 0.1850 = 0.3209$
$P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.3209 = 0.6791$

9. The Exponential Distribution

Suppose an event occurs with an average frequency of λ occurrences per minute and the average frequency is constant in that the probability that the event will occur during any tiny duration interval *t* is λt. Suppose further that we arrive at the scene at any given time and wait to the event occurs. Then the process can be described by an *exponential function*.

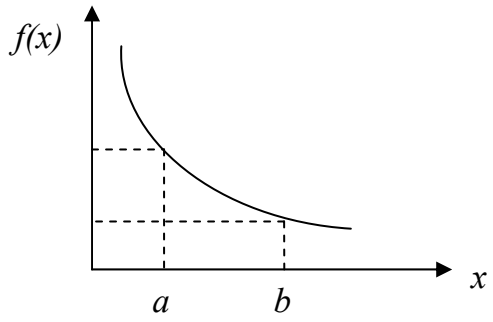
When *X* is exponential distributed with a frequency λ, we shall write $X \approx E(\lambda)$. The probability density function *f(x)* of the exponential function has the form

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0$$

Where λ is the frequency with the event occurs. The mean and the variance is given as

$$E(X) = \mu = 1/\lambda \qquad \text{and} \qquad V(X) = (1/\lambda)^2$$

Illustration



Exponential distribution formulas

$$P(X < a) = 1 - e^{-\lambda a}$$
$$P(X > b) = e^{-\lambda b}$$
$$P(a < X < b) = e^{-\lambda a} - e^{-\lambda b}$$

Example
Assume that the breakdown for a computer unit occur at $\lambda=2$ hourly intervals, then:

Mean:	$E(X) = \mu = 1/\lambda$	= ½ hours	(every 30 minutes)
Variance:	$V(X) = (1/\lambda)^2 = (½)^2$	= ¼ hours	(every 15 minutes)

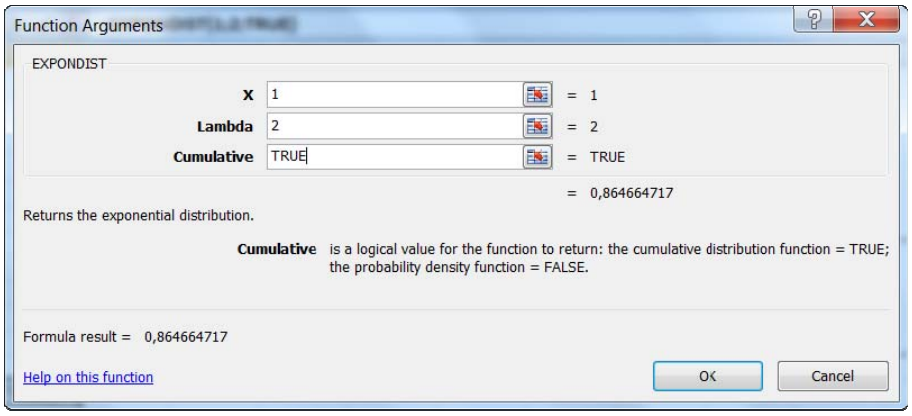
Then calculate for example:

$$P(X > 1) = e^{-2(1)} = 0.1353$$
$$P(X < 1) = 1 - e^{-2(1)} = 0.8667$$
$$P(1 < X < 2) = e^{-2(1)} - e^{-2(2)} = 0.1353 - 0.0183 = 0.1170$$

$P(X = 3) = 2e^{-2(3)} = 0.0049$ by direct use of the density function

Exponential Distribution in Excel
We can find the exponential distribution in Excel either directly or by approximation of an exponential function.

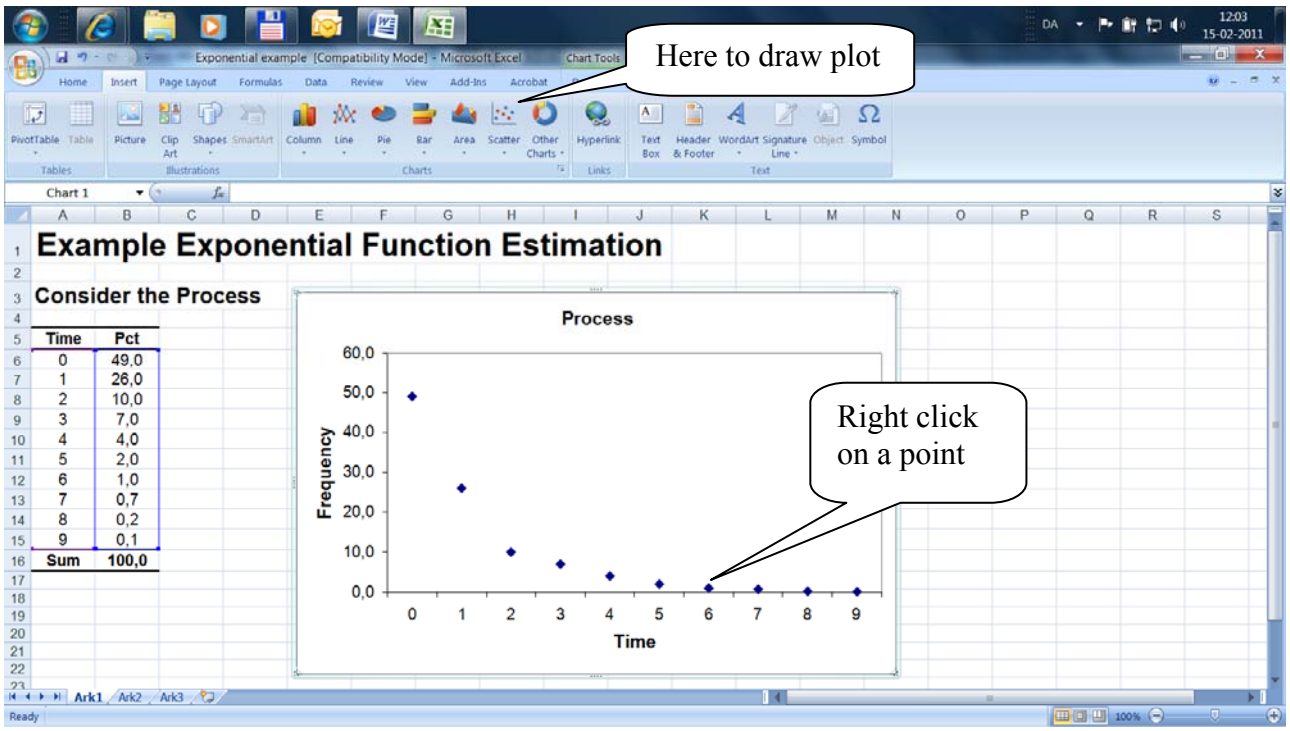
First try ***formulas/insert function/statistical/expondist*** and obtain the following screenshot



Here, I have solved the example for P(X<1) = 0.8647 in the little application just given. The difference with the result above on the third and fourth decimal is due to rounding off.

It is possible to use the scatter plot facility in Excel to approximate a data series to the exponential distribution and estimate the value of λ. This task is undertaken below. We consider an example of a data set with "queuing time" and "per cent" (of queuing). Do the following sequence:

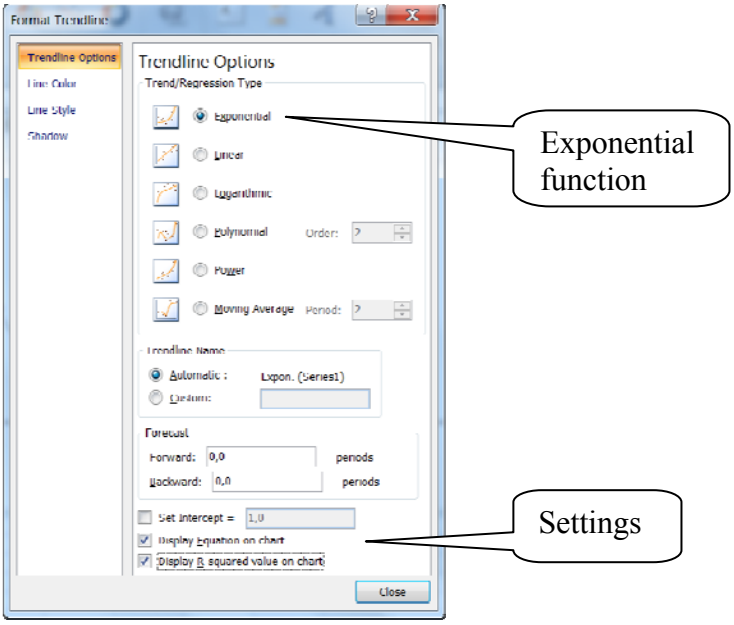
1. First construct a scatter (x-y) plot from the two columns the data set below. This should result in the plot displayed in the screenshot next page.



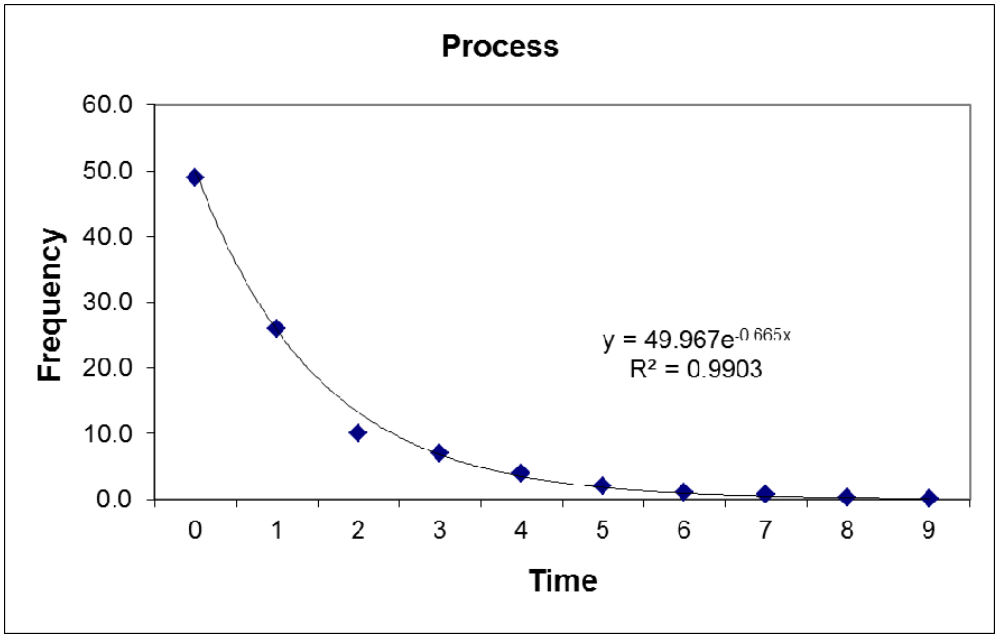
Then undertake the following sequence:

2. ”Right Click” on a point in the diagram
3. Click on the menu on ”add trend line”
4. On ”settings” remember to click on ”show equation”.
5. Mark “exponential equation”.
6. Click “ok” and the function is estimated.

These take are summarized in the following screenshots. First the ***add trend line menu***



7. The model is now estimated! It is observed from the final graph that $\lambda = 0.665$ in this case.



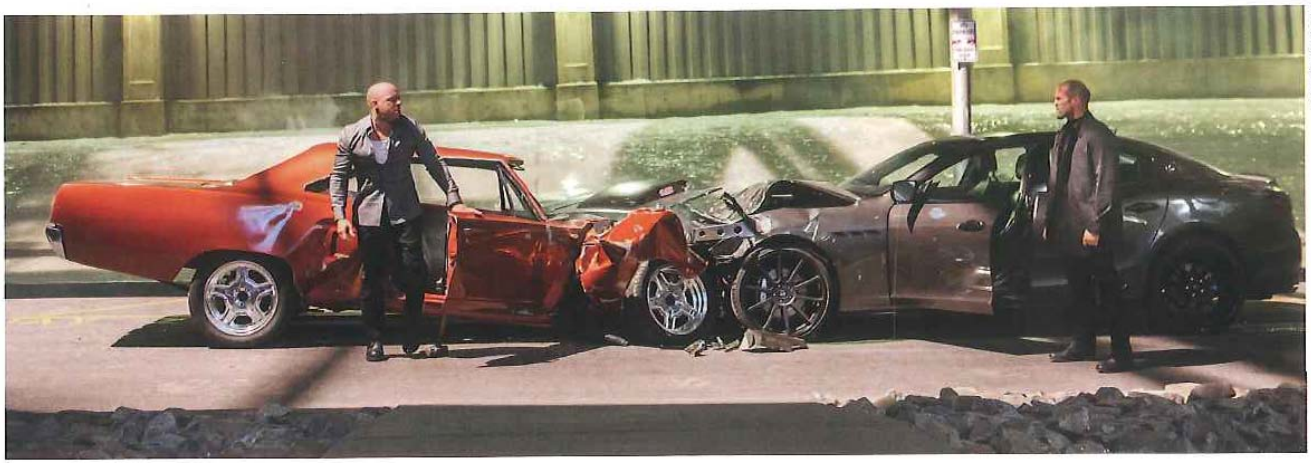
Example: Modelling Data from the Movie Sequel Fast and Furious

The exponential function is very efficient to model processes with an accelerating behavior like for example hyperinflation etc. In order to illustrate this feature consider an example.

The movie *Fast and Furious* was released in 2001. The style of the movie is as illustrated below! It is about fast cars, parties, car race and cozy ladies. In sum: These movies have everything that interests all males age 1 to 110!

With relatively fair reviews for the initial movie, and about 37 million tickets sold worldwide the set up called for a sequel. The followers have although of mixed reviews all been successes with lots of tickets sold.

Nearly nothing is left from the first movie. Actor Paul Walker died in a traffic accident in 2013. The only constant is producer H. Monritz and his Original Film Company.



The table summarizes the performance of the movies:

Some Statistics from the "Fast and Furious" Movie Sequels

Title	Year	No.	Spectators DK, 1000	Spectators World, mill	Positive reviews[1] %	Budget, mill USD
The fast and the furious	2001	1	50.852	36.7	53	38
2 Fast 2 Furious	2003	2	117.180	39.2	36	76
Fast & Furious: Tokyo Drift	2006	3	71.620	24.2	37	85
Fast & Furious:	2009	4	160.590	48.4	28	85
Fast & Furious 5	2011	5	187.590	79.0	78	125
Fast & Furious 6	2013	6	240.437	97.0	69	160
Furious 7	2015	7	371.250	184.9	81	190

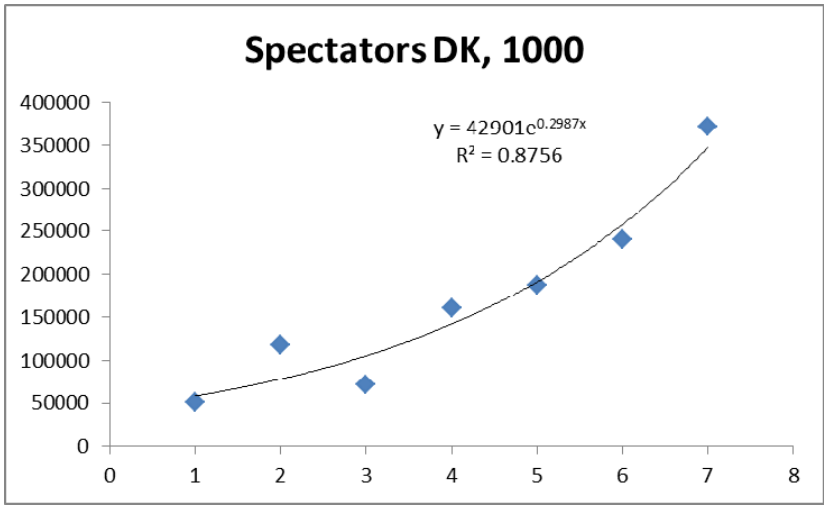
Note: 1) According to the portal "Rotten tomatoes".
Source: Statistics Denmark, IMDb, Box Office Mojo and US National Association of Theatre Owners.

The seventh movie gives a good idea of the future sequels. More extreme stunts and the main characters' have all supernatural powers.

The 8th fast and furious movie is planned to have premiere in the Danish cinemas on April 13th 2017. I'll properly be around!

The material in the table can be used for several purposes. Here I will attempt to predict the number of spectators expected in Danish theaters. To undertake this task I have edited a little in the table below, and created the table and graph below.

No.	Spectators DK
1	50852
2	117180
3	71620
4	160590
5	187590
6	240437
7	371250



I have provided a scatter diagram with the sequel number on the horizontal axis and the number of tickets sold on the vertical axis. As observed this is a non–linear function. As a result, the "add-trendline" function has been applied. The resulting exponential function is displayed in the graph.

The exponential relation has the form: $y = 42901e^{0.2987x}$

Substitute for x=8 and obtain: $y = 42901e^{0.2987(8)} = 468012$

When we get to the beginning of 2017 and the correct number is published, I will examine my result and evaluate on the validity of my prediction.

Nils Karl Sørensen ©
Statistics I, BA-INT

Set 4: Estimation and Confidence Intervals

by Nils Karl Sørensen

Outline *page*

Sections marked with an * will not be subject to independent exam questions.

1. Sampling and Point Estimation

What happens when we are sampling from a total population? Let us state some relevant questions:

- If the total population is Normal distributed, will the sample then be Normal distributed?
- What happens with the sample mean and the sample standard deviation? Will they be as in the total population?
- What if the total population is non–normal distributed? Will the sample be Normal distributed?
- If we take a large sample, will this sample be better than a small sample?
- What are the requirements of sampling?

How NOT to Sample

Let us begin our journey into sampling properties with the one of the most famous example on how NOT to sample. This must be the incorrect prediction of the outcome of the 1936 US presidential election! Today all elections are followed by for example exits-pools giving often quite good predictions on the outcome of an election from the minute the pool station close down.

In 1936 thing were different! Here the at that time widely quoted *Literary Digest* embarked on the project of prediction the outcome of the presidential election October 1936 where Republican A. M. Landon challenged the Democrat president F. D. Roosevelt.

The *Literary Digest* tried to gather a sample of staggering proportion – 10 million voters! One problem with the survey was that only a fraction of people sampled, 2.3 million, actually provided the requested information. Further, the sample of voters chosen by the *Literary Digest* was obtained from lists of telephone numbers, automobile registrations, and names of *Literary Digest* readers. However, in 1936 many people did not own a telephone or an automobile. The selection procedure of the sample was thus biased (slanted toward one kind of voters) because the sample was not randomly selected from the total population of voters.

As a result, well-heeled voters where over represented in the sample and these voters preferred A. M. Landon instead of F. D. Roosevelt, and his "New Deal"-policy[1]. The "New Deal" was a classical Keynesian expensive public policy stimulating public investment, reducing unemployment, and raising the tax rates for high-income earners. At that time, this was a very progressive public policy. In fact, the now classical book *General Theory* by J. M. Keynes had just been published a few months before the election took place.

[1] This policy has been similar to many of the packages implemented after the financial breakdown in 2007.

The Digest poll gave Landon 32 states and 370 electoral votes against 161 for Roosevelt. The result was a plurality to Roosevelt of approximately 11 million voters – history's largest pool! Roosevelt won 523 voters in the Electoral College to 8 won by Landon (he won Maine and Vermont). The *Literary Digest* lost so many readers that it went bankrupt! A famous photograph shows Roosevelt with a big smile showing up the front page of a newspaper with the misleading news.

By use of a small sample of only 2.300 voters, an unknown researcher named *Gallup* predicted the result correctly.

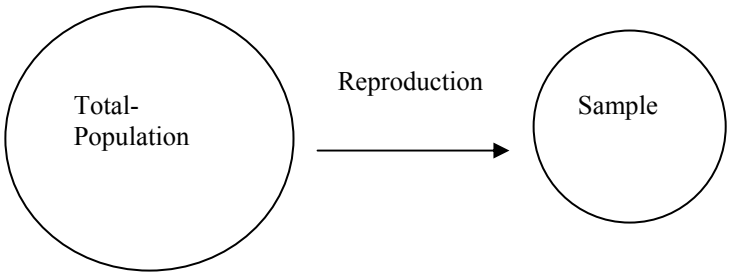
This is not the only time the US presidential election has experienced incorrect predictions. The picture shows a situation from the 1947 US presidential election where the winner H. D. Truman holds the first edition of the Chicago Daily Tribune with a front page that incorrect state him as the looser of the election to T. E. Dewey.

Estimators and Their Properties

What is the message? Well, a sample should be the *best* as possible, not as large as possible! The purpose of taking the pool is to provide a consistent *estimate* of the outcome of the election with regard to the position (the mean) and the dispersion (the standard deviation).

After the election, we have all information. From the total population we can find the outcome of the election. This is called the *population parameter.* When we take the sample we reproduce the total population as best as possible. From the sample we can estimate the

sample estimators. We can then compare the estimates of the population and the sample. In schematic form the process can be illustrated as:



That is that the sample mean and variance should be as close to the mean and variance of the total population as possible. Let us define:

- An *estimator* is an indicator for a *population parameter.* We can have either *point estimators* or *interval estimators.*

Let us introduce the following notation:

	Estimator (sample statistic)		***Population parameter***
Mean:	$\overline{X} \rightarrow$	estimates	μ
Variance:	$S^2 \rightarrow$	estimates	σ^2
Proportion share:	$\hat{p} \rightarrow$	estimates	p

We have earlier in the notes on *descriptive statistics*, defined the mean and the variance. The *sample proportion* is

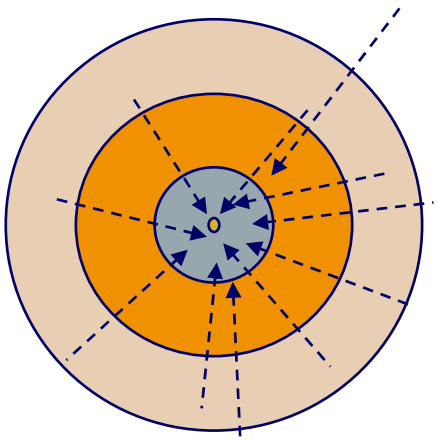
$$\hat{p} = \frac{x}{n}$$

Where x is the number of elements in the sample found to belong to the category of interest and n is the sample size.

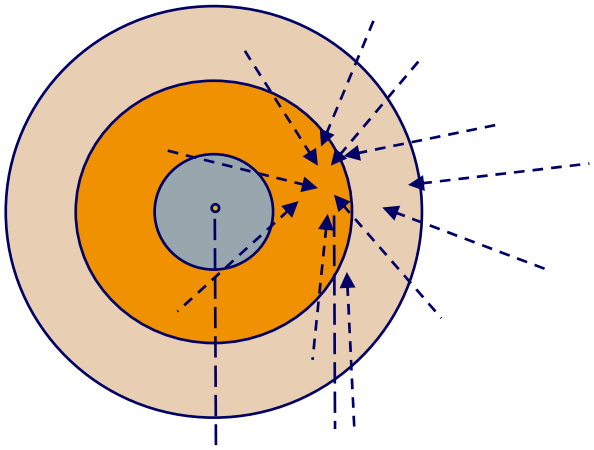
Sampling can be done with either **point estimates** or with **interval estimates** (this is mostly relevant in marketing). In the present case, we shall primarily focus on point estimates.

Rules and Criteria for Finding Estimates
Let us shortly lists the ***properties*** of a good estimator:

1. An estimator is said to be ***unbiased*** if its expected value is equal to the population parameter it estimates.



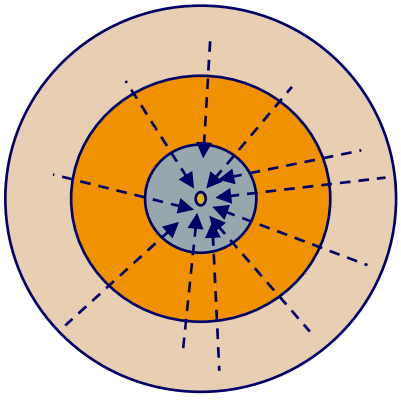
An **unbiased** estimator is on target on average.



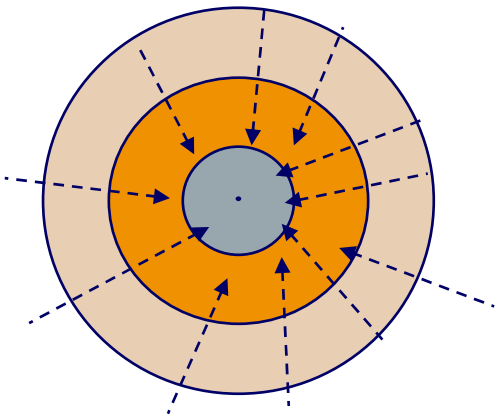
Bias

A **biased** estimator is off target on average.

2. An estimator is said to be ***efficient*** if it has a relative small *variance* (and standard deviation).

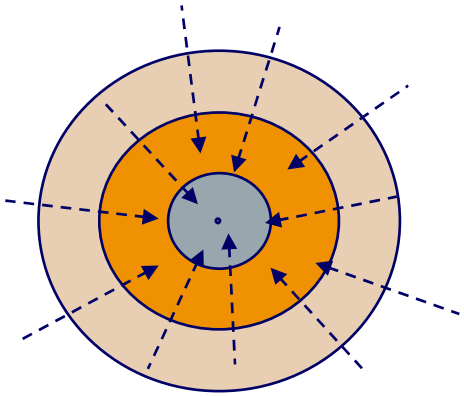


An **efficient** estimator is, on average, closer to the parameter being estimated.

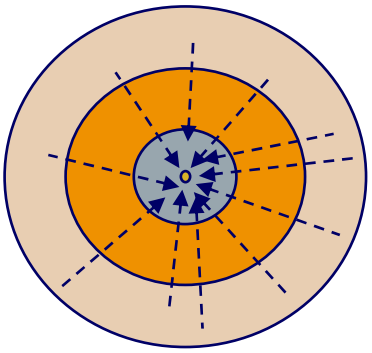


An **inefficient** estimator is, on average, farther from the parameter being estimated.

3. An estimator is said to be ***consistent*** if its probability of being close to the parameter it estimates increases as the sample size increases.



n = 10



n = 100

4. Finally, an estimator is said to be ***sufficient*** if it contains all the information in the data about the parameter it estimates.

Degrees of Freedom*

Back in the notes on *descriptive statistics*, we calculated the *population variance* σ^2 as well as the *sample variance* s^2. The latter, I divided by *(n–1)* rather than *n*. Here I will try to provide an explanation on this issue.

Suppose you are asked to choose 10 numbers. You then have the freedom to choose 10 numbers as you please, and we say that you have 10 *degrees of freedom*. However, suppose a condition is imposed on the numbers. The condition is that that the sum of all the numbers that you choose must be 100. In this case, you cannot choose all 10 numbers as you please. After you have chosen the ninth number, let us say that the sum of the nine numbers is 94. Your tenth number then has to be six, and you have no choice. Therefore, you only have *9 degrees of freedom.* In general, if you have to choose *n* numbers and a condition is imposed, you will only have *(n–1)* degrees of freedom.

Earlier the formula for the sample variance was stated as: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$ *or* $\frac{SSD}{n-1}$

where *SSD* is the sum of the squared deviations from the mean. Notice, that in the calculation of *SSD*, the deviations are taken from the sample mean $\bar{X}$, and not from the population mean μ. The reason is simple: While sampling, almost always, the population mean μ is *not* known. Therefore, we take the deviations from the sample mean $\bar{X}$. This will introduce some kind of bias. This is so because only by coincidence will $\bar{X}$ be equal to μ.

This bias is a restriction – just as in the example above where the ten numbers has to sum to 100. If $\overline{X} = \mu$ everything would be perfect. The sum of all observations divided by *n* (the sample size) would be equal to μ. The restriction is equal to using *(n–1)* instead of *n*. The sample variance s^2 will then be a little larger than the population variance σ^2 in order to capture the bias. We shall return to the impact of this feature, when we calculate the *confidence interval for the mean* later in these notes.

Let us consider a little experiment in order to explore what happens when we work with restrictions. Consider a sample of size six:

Sample	31	33	38	34	39	44

From this sample, we can calculate the sample variance. The sample mean is then $\overline{X} = (31+33+38+34+39+44)/6 = 219/6 = 36.5$. We can calculate the sum of the squared deviations from the left side of the table below. The variance is then found by dividing by (*n–1).* In this case the *degrees of freedom is (6–1)=5.*

We now chose two numbers from which deviations are to be taken namely for the first three sample points and for the last three sample points. These are given by $\overline{X}_1 = (31+33+38)/3 = 34$ and $\overline{X}_2 = (34+39+44)/3 = 39$. The right side of the table shows the new restricted calculation of the variance. In this case the *degrees of freedom is (6–2)=4.*

Sample	*Mean* $\overline{X}$	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	***Sample***	*Mean* $\overline{X}$	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$
31	36.5	–5.5	30.25	31	34	–3	9
33	36.5	–3.5	12.25	33	34	–1	1
38	36.5	1.5	2.25	38	34	4	16
34	36.5	–2.5	6.25	34	39	–5	25
39	36.5	2.5	6.25	39	39	0	0
44	36.5	7.5	56.25	44	39	4	16
		SSD =	**113.50**			**SSD =**	**67**

Calculation of the variance then results:

$$s_1^2 = \frac{113.50}{6-1} = 22.70 \qquad\qquad s_2^2 = \frac{67}{6-2} = 16.75$$

In case two, we use more information in order to obtain a more unbiased (and consequently lower) estimator of the variance. We use an extra degree of freedom in order to divide our sample into two parts. However, this also complicates matters because the information left for additional analysis will be limited.

We can now summarize how the number of degrees of freedom is determined. If we take a sample of size *n* and take deviations from the (known) population mean, then the deviations, and therefore the *SSD*, will have *df = n.* But if we take the deviations from the sample mean, then the deviations, and therefore the *SSD*, will have *df = n–1.* If we are allowed to take the

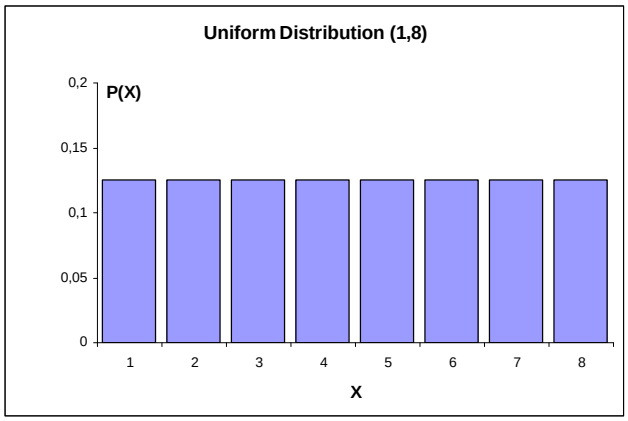
deviations from k ($\leq n$) different numbers that we choose, then the deviations, and therefore the *SSD*, will have $df = n–k$. While choosing each of the k numbers, we should choose the mean of the sample points to which that number applies.

The Central Limit Theorem

What happens to the estimators of the population parameters when an efficient sample is taken? The answer to this question is provided by the Central Limit Theorem.

Let us consider an example. In this case, we sample from a non–normal distribution and look at the outcome. We compare the mean and the variance of all samples with the total population.

Consider a uniform distribution of integers ranging from 1 to 8. We can illustrate the case as follows



Let us from this distribution calculate the expected mean μ, variance σ^2, and standard deviation σ. Let us in order to obtain consistence use the latter method or notation. Then obtain

X	P(X)	XP(X)	(X−μ)	$(X-\mu)^2$	$P(X)(X-\mu)^2$
1	0.125	0.125	−3.5	12.25	1.53125
2	0.125	0.250	−2.5	6.25	0.78125
3	0.125	0.375	−1.5	2.25	0.28125
4	0.125	0.500	−0.5	0.25	0.03125
5	0.125	0.625	0.5	0.25	0.03125
6	0.125	0.750	1.5	2.25	0.28125
7	0.125	0.875	2.5	6.25	0.78125
8	0.125	1.000	3.5	12.25	1.53125
Sum	1.000	4.500			5.25000

Then we obtain for the total population that

$$E(X) = \mu = 4.5 \qquad V(X) = \sigma^2 = 5.25 \qquad SD(X) = \sigma = 2.2913$$

Let us derive the *sampling distribution* of $\overline{X}$ in the case of drawing a sample of size ***n = 2*** items from the uniformly distributed population just given. This task can be undertaken in 8×8 = 64 different ways. If we first draw 1, then we can next draw 1,2, … , 8 etc. We can illustrate the possible samples outcomes of 2 items as

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)	(3,7)	(3,8)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)	(4,7)	(4,8)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)	(5,7)	(5,8)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)	(6,7)	(6,8)
7	(7,1)	(7,2)	(7,3)	(7,4)	(7,5)	(7,6)	(7,7)	(7,8)
8	(8,1)	(8,2)	(8,3)	(8,4)	(8,5)	(8,6)	(8,7)	(8,8)

Each of the samples has a mean, For example, the mean of the sample (1,4) is (1+4)/2 = 2.5, and the mean of the sample (8,4) is (8+4)/2 = 6.0 etc., Let us calculate all the possible sample means and obtain

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
2	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
3	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5
4	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5
6	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0
7	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
8	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0

The probability distribution of the sample is called the *sampling distribution of the mean.* Let us calculate this based on the 64 possible means just obtained. We get

X	P(X)	XP(X)	(X-μ)	$(X-\mu)^2$	$P(X)(X-\mu)^2$
1.0	0.0156	0.0156	-3.5	12.25	0.1914
1.5	0.0313	0.0469	-3.0	9.00	0.2813
2.0	0.0469	0.0938	-2.5	6.25	0.2930
2.5	0.0625	0.1563	-2.0	4.00	0.2500
3.0	0.0781	0.2344	-1.5	2.25	0.1758
3.5	0.0938	0.3281	-1.0	1.00	0.0938
4.0	0.1094	0.4375	-0.5	0.25	0.0273
4.5	0.1250	0.5625	0.0	0.00	0.0000
5.0	0.1094	0.5469	0.5	0.25	0.0273
5.5	0.0938	0.5156	1.0	1.00	0.0938
6.0	0.0781	0.4688	1.5	2.25	0.1758
6.5	0.0625	0.4063	2.0	4.00	0.2500
7.0	0.0469	0.3281	2.5	6.25	0.2930
7.5	0.0313	0.2344	3.0	9.00	0.2813
8.0	0.0156	0.1250	3.5	12.25	0.1914
Sum	1.0000	4.5000			2.6250

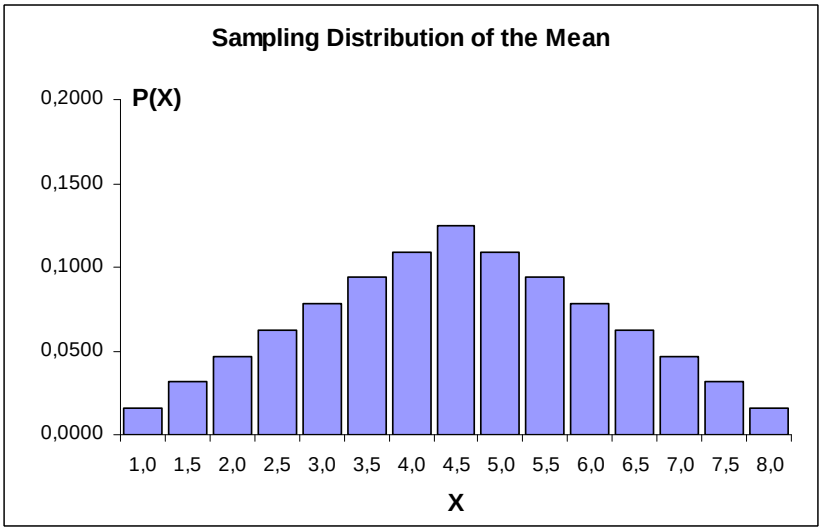
Then we obtain from the samples that

$$E(\overline{X}) = \mu = 4.5 \qquad V(\overline{X}) = \sigma^2_{\overline{X}} = 2.6250 \qquad SD(\overline{X}) = \sigma_{\overline{X}} = 1.6202$$

This is surprising! We obtain *a similar mean*, but *the variance is* ***smaller***. To be exactly, the variance is exactly half of what we expected. However, the sample size was equal to $2 = n$. This imply that

$$E(\overline{X}) = \mu \qquad V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2 / n \qquad SD(\overline{X}) = \sigma_{\overline{X}} = \sigma / \sqrt{n}$$

Let us look at the distribution of the means by plotting X and $P(X)$ from the table above:



This ”pyramid” look like a normal distribution. Let us summarize out findings:

When sampling from a Uniform Distribution (non–normal distribution) then the mean will be Normal Distributed and:

$$X \approx U(a,b) \rightarrow \overline{X} \approx N(\mu, \frac{\sigma^2}{n})$$

Let us provide a generalization. Thus, when we sample from any distribution with mean μ, and standard deviation σ, the sample mean has a normal distribution with the same *center* μ as the population but with the w*idth* (standard deviation), that is $1/\sqrt{n}$ the size of the width of the population distribution. Consequently, we have an *unbiased* estimator of μ.

This is illustrated below, with mappings of different distributions of $\overline{X}$ with different sample sizes. Notice that as n increases the width becomes smaller. Stated differently the standard deviation of the sample becomes smaller as the size of the sample increases. So in a large sample, the likelihood that the sample mean $\overline{X}$ will be close to the population μ increases. Our estimator is in general *consistent, unbiased, efficient* and *sufficient*. All four requirements are fulfilled. From the illustration, it is observed that as n increases all sampling distributions becomes normal and equally alike regardless of the initial type of distribution. We now obtain the *central limit theorem:*

The Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean $\overline{X}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size becomes large

For "large enough" n $\qquad \overline{X} \approx N(\mu, \sigma^2/n)$

We expand the central limit theorem to also be valid for the distribution of the sample proportion $\hat{p}$.

The Central Limit Theorem for Sampling Proportions

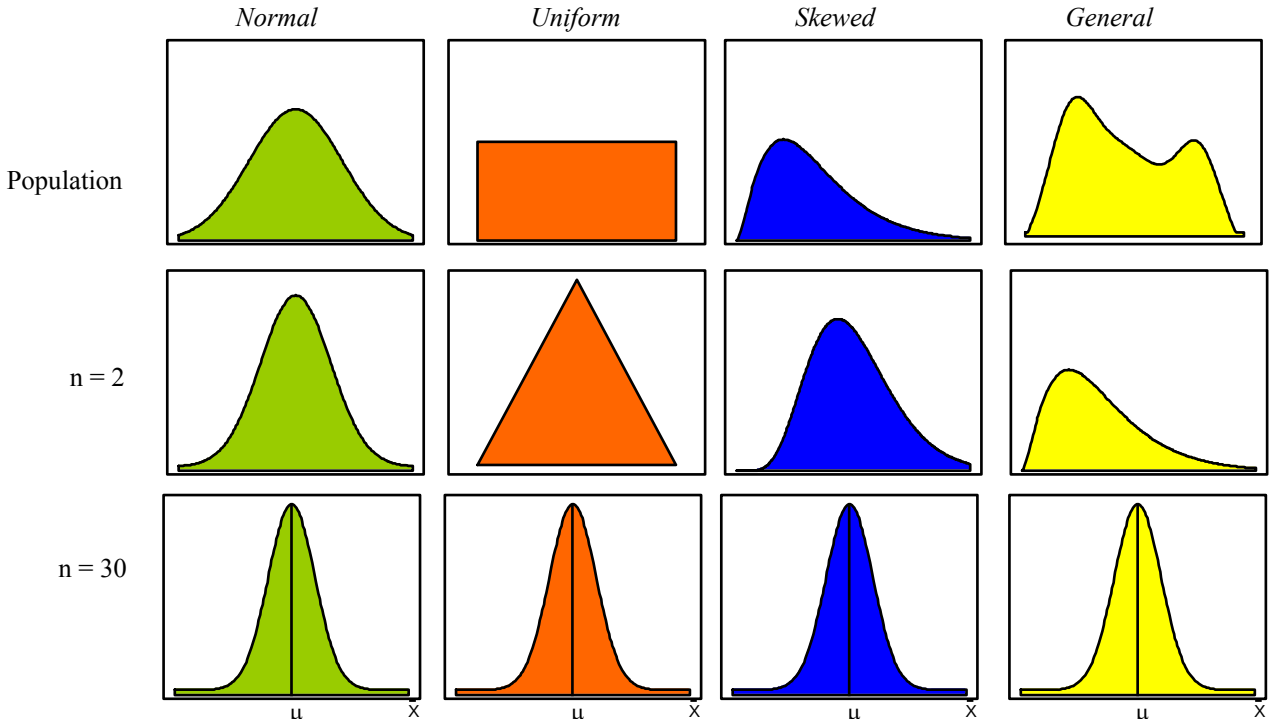
The population mean of all possible sample proportions $\hat{p}$ will be an unbiased estimator of the proportion parameter p with mean and variance

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

This is valid when the sample size n is large.

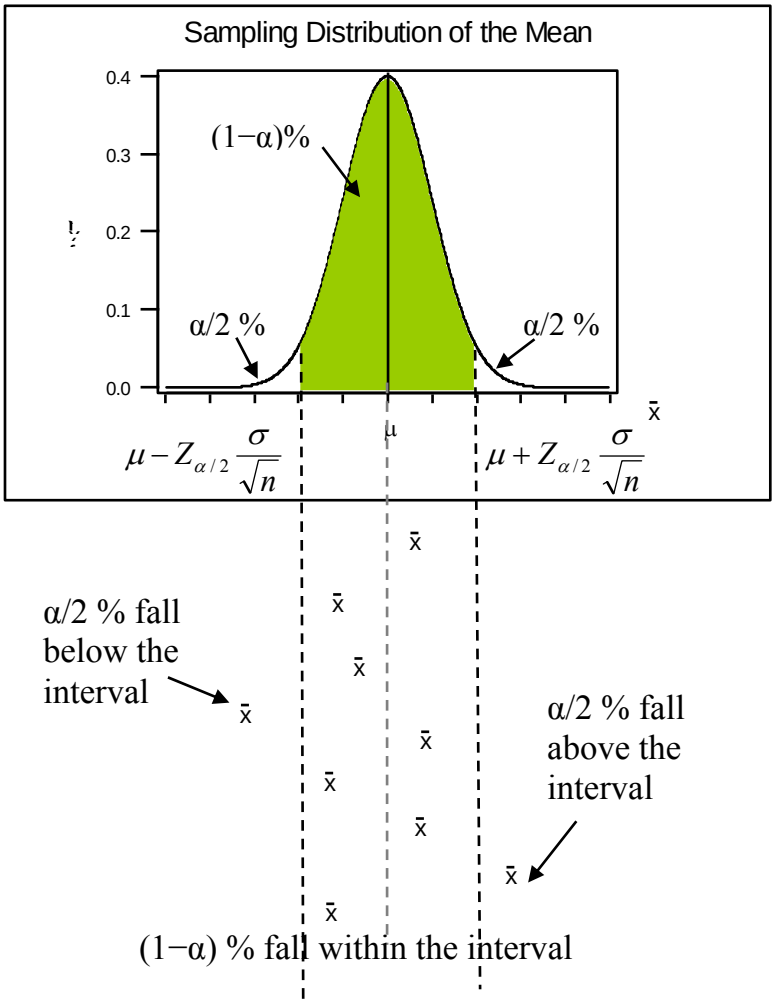
The sample size n should be considered as large if both np and $n(1-p)$ are at least 5.

The Effects of the Central Limit Theorem: The Distribution of $\overline{X}$ for Different Populations and Different Sample Sizes



2. Z-based Confidence Intervals for a Population Mean: σ Known

If we take many samples from a large total population, and take the mean of this, then a *confidence interval* is the range of numbers believed to include an unknown population parameter (μ or $\hat{p}$). Associated with the interval is a measure of the confidence. We have that the interval does contain the parameter of interest. Let us illustrate the situation:



The idea is illustrated above. *Before* sampling, we consider the normal distribution in the top of the figure. Then we – just as the case when we worked with the example on the uniform distribution earlier – take samples of the total population. The resulting samples each have a mean $\overline{X}$. Some of the means are more likely to occur than others. If our sample not is biased the sample mean $\overline{X}$ will be close to the population mean μ.

In the present case σ is known from the total population. Further, as the total population is normally distributed this must also be valid for the sample. If the central limit theorem is

valid, and the inverse transformation can be written as $\bar{X} \pm z\sigma$, then we can define a $(1-\alpha)$ 100% confidence interval for μ when σ is known and sampling is done from a normal distribution, or with a large sample as

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

So $(1-\alpha)$ tells us where we find the majority of the sample means. We use $\alpha/2$ because the interval is two-sided. The tolerances for the confidence intervals depend on the *confidence level.* The typical confidence level is *95 percent.* This means that $\alpha = 0.05$ and $(1-0.05)\times 100 = 95$ percent. As the interval is two-sided the lower bound will be found at 0.05/2 equal to 0.025, and the upper bound will be found at $(1 - \alpha/2)$ equal to 0.975.

The values of Z, corresponding to these probabilities can be found by use of the tabulation of the Normal distribution in **Statistics Tables**. It is only necessary to find one of the values of Z. This is so because the Normal distribution is symmetric around the mean that is equal to zero.

Z	**0.00**	**…**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
1.6	0.9452	…	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	…	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	…	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	…	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	…	0.9798	0.9803	0.9808	0.9812	0.9817

In the extract from the table of the Normal distribution above it is show how the upper bound is found to equal Z = 1.96. Find as an exercise the lower bound at the probability equal to 0.025. For a 95 percent confidence interval it is found that Z = ±1.96.

The commonly used confidence levels are either on 90, 95 or 99 percent. *If nothing is stated for example in an exam exercise the default is always the 95 percent level.* The corresponding values of Z for different confidence intervals are given as:

Confidence interval:	**Value of Z**
90 ($\alpha = 0.10$) 95 ($\alpha = 0.05$) 99 ($\alpha = 0.01$)	± 1.645 ± 1.960 ± 2.575

Confirm as an exercise the values of Z for the 90 and 99 percent confidence intervals.

Example
Assume a population size equal to *n = 100,* standard deviation equal to 5 and mean equal to 50. Set up a 90 %, a 95 % and a 99 % confidence interval for the sample mean.

To solve the problem we apply the formula, above and the z-values taken from the table above.

90 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 1.645\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 0.8225 \Rightarrow [49.1775\,;50.8225]$

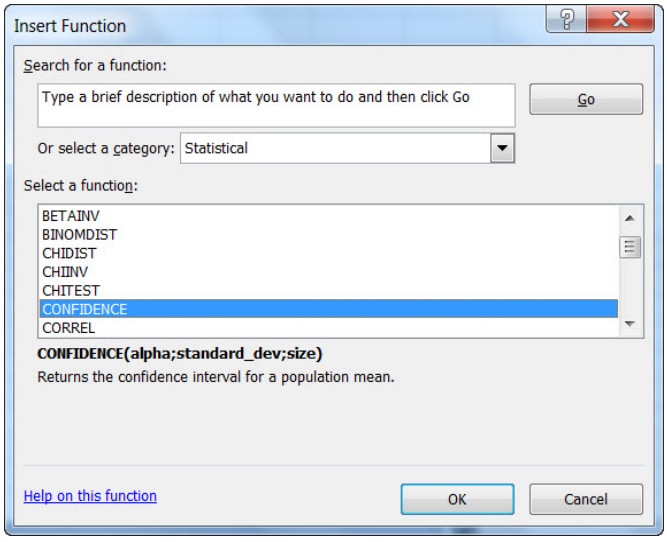
95 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 1.960\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 0.9800 \Rightarrow [49.0200\,;50.9800]$

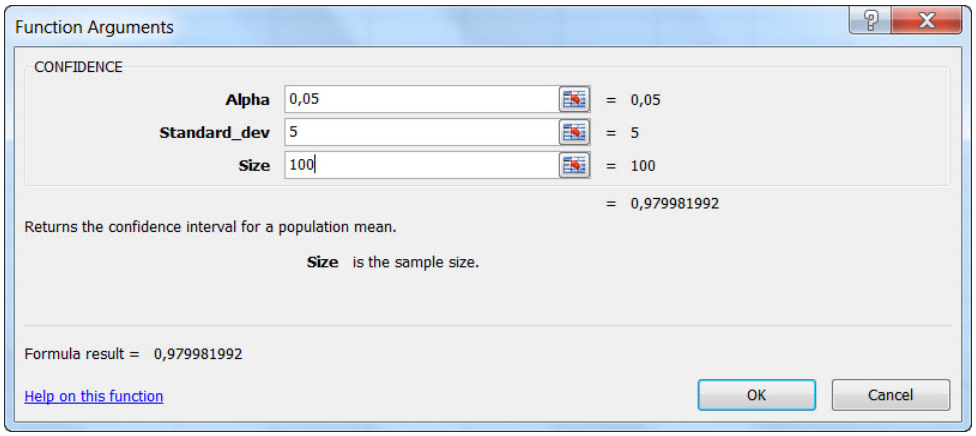
99 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 2.575\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 1.2875 \Rightarrow [48.7125\,;51.2875]$

Notice that the interval becomes larger as α decrease.

The confidence intervals can be found by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → 7: Zinterval. In the second line STATS should be marked. Now insert the values in the menu. In the present case σ = 5, $\overline{X}$ = 50 and n = 100. Select C-level: .95 → CALCULATE → ENTER. Then the result will appear as [49.02;50.98]. Just as calculated above.

In Excel, we can set up a confidence interval either in *descriptive statistics* or by use of **formulas | insert functions | statistical | confidence.** Then we obtain the picture below. Click on "OK" and obtain the picture on the next page.





Example of Exercise with Confidence Intervals on the Mean: The Case of Gasoline

Let us move back to the example of the price for gasoline also used in my notes in set 2 on the Normal distribution. Measured over 25 visits, the mean price of 95 octane gasoline was 9.95 DKK and the standard deviation was 0.30 DKK.

On this basis, answer the following question:

- Set up a 95 % confidence interval for the mean price of gasoline.

Using the information the confidence interval can be set up as:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 9.95 \pm Z_{0.025}\frac{0.30}{\sqrt{25}} \Rightarrow 9.95 \pm 1.96(0.06) \Rightarrow 9.95 \pm 0.1176 \Rightarrow [9.8324\,;10.0676]$$

The interpretation of the confidence interval is that in 95 percent of the visits to the gasoline station the price will vary according to the interval. That is a price ranging between 9.83 DKK and 10.07 DKK.

Example of Using the Confidence Interval to Examine the Validity of a Statement (exam July 2012, 10%)

Frequently the confidence interval is used to examine the validity of a statement. First, the confidence interval is calculated. The confidence interval is then compared with the statement. Finally, a conclusion is drawn. If the statement is outside the range of the confidence interval the statement is rejected as being false. If the statement is inside the range of the confidence interval the statement is accepted as being valid with the probability given by the confidence interval.

Case

The trip from Hamburg to Flensburg by train with the "Regionalexpress" takes precisely 120 minutes. A sample of 100 representative trips by car on the motorway resulted in a mean travelling time by car equal to 110 minutes with a standard deviation equal to 30 minutes. This variation is due to queues, low speed due to renovation and construction work etc. It is assumed that the data set is Normal distributed.

Is the trip by car significantly different from the trip undertaken by train?

Solution

First the statement has to be identified. This must be a comparison between the 120 minutes by train and the 110 minutes by car.

Although this trip looks to be less time consuming there is uncertainty due to road works, football at Volkspark or whatever.

To examine the issue a 95 percent confidence interval is set up for the trip by car. It is observed that $n = 100$ and in addition the standard deviation is given. The most important is the formulation of the exercise is that the data set is assumed to be *Normal distributed.* This is a hint saying that this exercise can be used by setting up a confidence interval for the travelling time by car.

By substitution and assuming a 95 percent level of significance a confidence interval can be set up.

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 110 \pm Z_{0.025}\frac{30}{\sqrt{100}} \Rightarrow 110 \pm 1.96(3) \Rightarrow 110 \pm 5.88 \Rightarrow [104.12\,;115.88]$$

It is observed that the 120 minutes of the train trip falls outside the interval. The implication is that the trip by car it the fastest.

(However, it is not very much faster, and you can also relax and take a nap in the train☺)

3. Working with Population Proportions

Sometimes interest centers on a qualitative, rather than a quantitative variable. We may be interested in the relative frequency of occurrence of some characteristic in a population. For example, we may be interested in the proportion of people in a population who are in favor of a given agenda. In such a case, we want to estimate the population proportion *p*.

An estimator of *p* is the sample proportion $\hat{p}$. Earlier in these notes, we saw by use of the Central limit Theorem that when the sample size is large $\hat{p}$ has an approximately normal distribution with mean $\hat{p}$ and variance $\sqrt{\hat{p}(1-\hat{p})/n}$. We can use this information to set up a confidence interval for the population proportion along the same lines as we did for the sample mean.

A large sample (1−α) 100% confidence interval for the population proportion *p* is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where the sample proportion $\hat{p}$ is equal to the number of successes in the sample *x*, divided by the number of trials (the sample size) *n*, and $\hat{q} = (1-\hat{p})$.

Here *n* should be considered large if both $n \cdot \hat{p}$ and $n \cdot (1-\hat{p})$ are at least 5.

Example: The Backer from Langeland and the Local Election
During the past decade, the number of municipalities in Denmark has decreased. For example on the island of Bornholm 5 municipalities were after a vote unionized into a single municipality.

On February 12th 2003 a similar agenda was considered by a vote for the three municipalities on the island of Langeland located south of Fuen.

A local baker provided up to the vote an interesting form of projection of the outcome of the vote. He sold a cookie called a Shrovetide bun (In Danish "Fastelavnsbolle"), and the person who board such a cookie, had the choice of having either the word "yes" or "no" decorated by chocolate on the cookie.

Up to the date of the vote the baker managed to sell a total of 385 cookies decorated with "yes" and "271" decorated with a "no".

- Set up a 95 % confidence interval for the proportion share of cookies with "yes".

The total number of sold cookies is equal to n = 385 + 271 = 656.

The proportion of "yes" is equal to $\hat{p} = x/n = 385/656 = 0.587$. Now we can use the formula just given and set $Z_{0.025} = 1.96$

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow 0.587 \pm 1.96\sqrt{\frac{0.587(1-0.587)}{656}} \Leftrightarrow 0.587 \pm 1.96(0.019) \Leftrightarrow 0.587 \pm 0.038$$

Then the interval is equal to $[0.549\,;\,0.625]$. The condition that $n \cdot \hat{p} \geq 5$ is met.

Example with an Investigation of a Statement (exam July 2012, 10 %)

Similar to the procedure outlined in the last section the validity of a statement for a population proportion can be examined by setting up a confidence interval.

Case

A fitness center in Flensburg has a market share equal to 18 percent. The center undertakes a market campaign. After the campaign a questionnaire is undertaken. Here 300 representative selected persons are asked if they want to use the fitness center. Out of the persons asked 25 percent responds with a "yes".

The management of the fitness center evaluates that the market share has changed significantly as a result of the market campaign. Can the statement put forward by the management be supported by the statistical evidence assuming that the data set is Normal distributed and a 95 percent level of confidence?

Solution

Initially the statement has to identified. Before the campaign the share was 18 percent, and now 25 percent out of the 300 persons asked claims that they will use the fitness center. So the 18 percent has to be compared by the 25 percent out of the 300 respondents.

This means that $n = 300$ and $\hat{p} = 0.25$.

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow 0.25 \pm 1.96\sqrt{\frac{0.25(1-0.25)}{300}} \Leftrightarrow 0.25 \pm 1.96(0.025) \Leftrightarrow 0.25 \pm 0.049$$

A 95 percent level confidence interval is then equal to $[0.201\,;\,0.299]$.

The assumptions $n \cdot \hat{p} \geq 5$ and $n \cdot (1-\hat{p}) \geq 5$ are both fulfilled.

By comparison it is observed that the 18 percent or 0.18 is found to be outside the range of the confidence interval. So in this case the statement is valid because the 25 percent is significantly different from the 18 percent.

Nils Karl Sørensen ©
Statistics I, BA-INT

Set 5: Scale Levels and Cross Tabulations

by Nils Karl Sørensen

Outline *page*

1. Scale Levels

So far we have worked with investigations of single variables. We have undertaken various calculations and worked with the mean, the standard deviation etc. In the present set of notes we shall look on the types of variables, and how these variables are measured. In addition, we shall examine how two set of variables can be tabulated by use of a cross tabulation.

A table consists of counting of numbers. It can be unemployed persons, cars, the number of students attending the course in Statistics I at the BA-INT education etc. The numbers that are counted and classified are labelled *elements.* For example there can be 110 elements or students in the class attending the course in Statistics I.

The elements can be *classified* into *categories* by certain *characteristics*. For example the 110 students can be divided by gender. So gender is the classification variable of the two characteristics or *criteria's.* There may for example be 60 females in the class and 50 males.

There are two types of *characteristics*; the direct measureable criteria's and the indirect measureable criteria's. Let us examine these by turn.

The direct measureable criteria's are called ***quantitative*** criteria's. In this case a numeric value can be associated. This can for example be prices measured in € or in DKK or data can be as an index relative to a base year. Other examples could be the measurement of the height of a person in centimeters or the weight of the person measured in kilos. In the course we have applied quantitative data to calculate measures of location and dispersion. We have used the computed values to identify the shape of the distribution of the considered data.

The indirect measurable criteria's are called ***qualitative*** criteria's. Numeric values cannot be associated with the data. Instead the characteristics are counted and measured in units. Examples of qualitative variables are gender, civil status, residence, level of education, position etc. For such data the calculation of the mean, median etc. will not give any meaning. However, a data set of for example the income can be divided by gender and then the mean income for females and males can be computed. The counting variable can take the values female=1 and male=0. In this case the qualitative variable provides information that can be used in the further analysis.

The quantitative and qualitative criteria's can each de decomposed into two groups depending on the classification. Then the following four classifications (measurement scales) appears:

Qualitative criteria's:	1) Nominal scale	
	2) Ordinal scale	
Quantitative critiria's:	3) Reference point scale	} Interval scale
	4) Zero point scale	

Consider the four scales by turn. First the qualitative criteria's.

***Ad 1) Nominal scale*:** Classification of elements by qualitative criteria's with out any given order.

The classification is only a sorting or dividing. The elements could for example be persons with residence in Denmark divided by gender, occupation or municipality. The order of the sorting has no influence on the outcome of the analysis.

Ad 2) Ordinal scale: In this case the order has an influence. It could for example be the classification of nations by credit ranking (AAA, AA+, etc.), or it could be levels of utility as known from microeconomics.

Ordinal scales are used very frequently for example in questionnaires. In questionnaires there will be a number of outcomes associated with a given question. Then the person or respondent being questioned can mark the preferred outcome/answer. For example the outcomes can be labelled as:

☐	☐	☐	☐	☐
Very satisfied	Satisfied	Neutral	Unsatisfied	Very unsatisfied

The answer to the question is a ranking. It is not possible to determine how large the difference is between AAA and AA+ or between ”satisfied” or ”neutral”. Futher, the judgement or valuation of the outcome may be individual. For example two credit rating companies may have two different credit ratings of a given country. Two respondents may also have two different judgements about what is “satisfied” or “neutral”.

It is possible to undertake some calcalations on data from a questionnaire given in the from outlined by the formulation of the outcomes above. For example diffenrent numeric values could be associated with the different possible answers. This could look as:

☐	☐	☐	☐	☐
Very satisfied=1	Satisfied=2	Neutral=3	Unsatisfied=4	Very unsatisfied=5

Now it is possible to conduct numerical analysis on the material. As an example, assume that there are 200 respondents who answered a given question with the outcomes as given above. A table with the frequency distribution by outcome could be as follows:

Answer	Very satisfied	Satisfied	Neutral	Unsatisfied	Very unsatisfied	
Number	1	2	3	4	5	
Frequency	20	40	80	30	30	200

Now it is possible to calculate for example the mean and other measures of location or dispersion. For example the mean may be equal to 3.05 and the standard deviation may

equal 1.16. What is the interpretation? In this example the mean answer is very close to "neutral". Further most of the answers will fall within the range from 2 to 4. This is evident from the low standard deviation. In addition to the descriptive statistical analysis the distribution of the frequencies of the answers can be displayed by a histogram. However, observe that out data are normative. So it is not possible to give the interpretation that "satisfied" is one unit better than "neutral".

Consider a different question from the questionnaire. Here the distribution of the frequencies can be observed from the table

Answer	Agree much	Agree	Neutral	Disagree	Disagree much	
Number	1	2	3	4	5	
Frequency	60	30	20	40	50	200

Assume that the mean here is equal to 4.20 and the standard deviation is equal to 3. Is it possible to compare this outcome with the outcome from the previous table above? The answer to this question is no. This is so because the classifications of the outcomes not are similar. For example "satisfied" may not be the same as "agree". In order to provide a comparison the outcomes must be similar. The conclusion is that the ordinal scale is complicated to work with.

Having discussed the qualitative criteria' consider next the quantitative criteria's. Contrary to the example above the measurement scale can be associated with an interpretation. An ***interval scale*** is a measurement scale for quantitative data. An interval scale can take two forms. It can either be a point scale or an interval scale. In descriptive statistics we have worked with both types of scales. A *point scale* could for example be a times series at points 1, 2, 3, etc. whereas an interval scale uses data divided by segments i.e. intervals for example [0−9] ; [10−19] etc.

Notice that the point of departure is different in the two examples. The former scale starts in the point 1, whereas the latter includes 0. What is correct? In order to examine this issue consider the two classifications of the interval scale.

Ad 3) Reference point scale: Classification of elements after quantitative criteria's and divided by an interval scale with arbitrary chosen reference point.

Consider for example the Celsius and the Fahrenheit scales of temperature. For the Celsius scale the reference point is 0, but at this temperature the Fahrenheit scale is equal to 32. The time line is another example. In the Christian culture the year of reference is year 0, whereas the time line in Japan referees to the year of the sitting emperor. In the Muslin culture the

base year 0 is year 622 after Christ. Other examples are the grade scales for example the Danish and the German grade scales.

Ad 4) Zero point scale: Classification of elements by quantitative criteria's and divided by an interval scale with an unambiguous reference point. This could be height, weight, distance, diameter, space, prices, earnings, costs, income, savings, consumption, age or speed. These are all examples where the measurement is determined by use of a zero point scale.

Example

Problem 7, Summer Exam 2011 (3 points, 15 %)

A biological investigation has examined the number of butterfly species in different biotopes in the German state of Schleswig-Holstein. The investigation considered various areas of forest and areas of meadows. The table records the number the number of butterfly species observed by area or biotope.

Area no.	Meadows
	Number of species
1	8
2	11
3	10
4	10
5	8
6	14
7	11
8	14
9	11

Area no.	Forest
	Number of species
1	8
2	10
3	5
4	5
5	5
6	12
7	17
8	5
9	8

A total of 18 areas were investigated. Out of them 9 were meadow areas and 9 were areas coved with forest.

***A)* 0.5P**

What scale level is being used?

The scale unit is "number of butterfly species observed" divided by biotope (area type). Data are quantitative by areas. Here a point scale is used. The biotope is a nominal scale variable just as for example gender.

***B)* 1.5P**
Calculate all possible measures of location i.e. the mean, the median and the mode for the number of butterfly species in both areas or biotopes.

This task is done by use of Excel or by use of a **pocket calculator.** The upper quartile and lower quartile has also been calculated. The material is summarized in the table.

Measure of location	Meadows	Forrest
	Number of species of butterfly	
Mean	10.78	8.33
Mode	11	5
Median	11	8
Q_1	9	5
Q_3	12.5	11

The mode has been found by inspection of the table with the data.

***C)* 0.5P**
What type of distribution of data is observed? Motivate your answer.

For *meadows* the mean, the median and the mode are very identical. Here a *symmetric distribution is observed*. With regard to the *forest* the mode is lower than the median and the mean. This distribution is then *symmetric to the right*.

***D)* 0.5P**
If you have to go on a Sunday trip to watch butterflies where will you go? Motivate your answer.

It is the best to go to the meadows. Here the number of species of butterflies is the highest. Further, with the symmetric distribution it is more likely that the mean number of species will be observed.

2. Cross Tabulations (Pivot tables)

The table is one of the most important tools in practical statistics. The table provides a systematic and detailed presentation of data. The table also serves as a point of departure for the typical empirical investigation or report.

Clearness is the keyword for a table. Here unorganized amounts of data can be displayed in an organized way; counted and presented. The main structure of a table is as follows:

	Main or head column			
Front column				

The front column has normally only a single column, whereas the head column may have several columns. Let us try as an example to set up a table for persons by specific intervals of income for the year 2010.

	Persons by income (1.000 DKK), persons			
	0 – 99	100 – 199	200 – 399	400 or more
2010				

In addition, the table needs to have a headline. The headline should be short, but it should also cover the content being presented in the table. There should also be a table number. Under the table the source should be located along with eventually notes. Here for example data beaks, changes in definitions etc. should be described.

It should be considered if the table provides a reasonable description of the topic being analyzed. What kind of data is contained in the table? Maybe a percentage distribution of data is desireable? Different possibilities should be testes before a final decision is taken.

Let us expand the table. For example it may be relevant to consider the data divied by gender. Then the table could look as:

2010	Persons by income (1.000 DKK), persons			
	0 – 99	100 – 199	200 – 399	400 or more
Female Male				

If two elements are combined in a single table then it table is called a *cross table* or an *Pivot table.* In the example gender and income are combined.

Let an example illustrate the use of cross tables. By use of a questionnaire 77 students at the SDU campus Sønderborg were asked about their preferences for Coca-Cola. It was possible to choose between ordinary Coca-Cola or diet Coca-Cola. At the questionnaire the students were also asked about their gender.

The answers were counted and combined into a cross table. There are two types of Coca-Cola and the two genders. In addition, the totals can be calculated. Undertaking this task results in a cross table looking as

	Cola	Diet cola	**Total**
Female	3	33	36
Male	30	11	41
Total	33	44	77

Inspection of the table reveals that the choice of Coca-Cola and gender are related. Out of the female students 33 prefers the diet Coca-Cola out of a total equal to 36. This is more than 90 %. With regard to the male students 30 prefers the standard Coca-Cola out of a total equal to 41. This is about 73 %. In this case *dependence* is observed among the gender and the choice of Coca-Cola. The females prefer the diet Coca-Cola whereas the males prefer the standard Coca-Cola.

How would this table look if dependence not were present? Then about half of the females as well as the males would prefer one of the types of Coca-Cola. So both genders would be indifferent about the choice of Coca-Cola. In the table below the data has been reshuffled such that *independence* is present.

	Cola	Diet cola	**Total**
Female	18	18	36
Male	20	21	41
Total	38	39	77

This is a table with the feature of *independence.* In this table gender and the choice of Coca-Cola are not related. Maybe other things may influence on the choice of Coca-Cola, but not gender. In the course of Statistics II we shall consider a more specific procedure to examine for independence.

Example

Problem 1 and 2, Summer Exam 2011
(both problems has been included because problem 1 serves as basis for problem 2)

Problem 1 Summer Exam 2011. (2 P, 10 %)
The table below displays the number of "exam point" obtained for 15 students in an exam in German language.

Point	10	14	7	18	12	15	12	12	8	8	4	9	21	11	15

***A)* 1P**
Set up a table of the relative and cumulative frequencies in order to display data by a relevant category of measurement. Info: The exam was passed with 10 points.

For the table it is for example observed that two students obtained 8 points. Counting the numbers results in the following table:

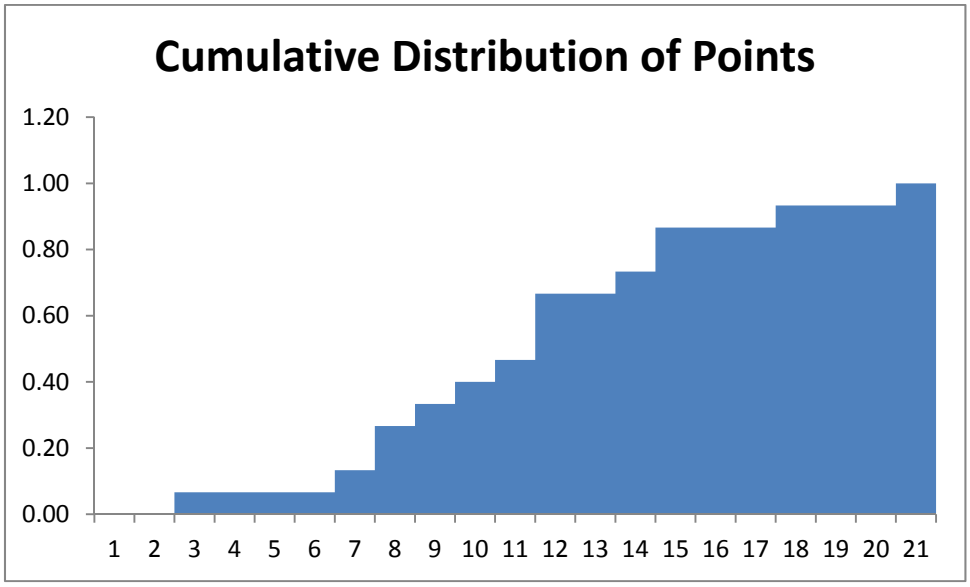
Points	4	7	8	9	10	11	12	14	15	18	21	Total
Counting	|	|	||	|	|	|	|||	|	||	|	|	
Frequency	1	1	2	1	1	1	3	1	2	1	1	15
Relative frequency	0.07	0.07	0.14	0.07	0.07	0.07	0.21	0.07	0.14	0.07	0.07	1.00
Cumulative frequency	0.07	0.14	0.28	0.35	0.42	0.49	0.70	0.77	0.91	0.98	1.00	(rounded off)

The table can be reduced by use of the information that 10 points are needed in order to pass the course. The table then looks as

	Not passed	Passed	Total
Students	5	10	15
Relative frequency	0.33	0.67	1.00

***B)* 1P**
Provide a graphic illustration of the data for "exam points" as detailed as possible.

Using the cumulative data form the table the graph looks as:



Problem 2 Summer Exam 2011 (3 P, 15 %)
The table below adds an extra dimension to the data set on exam points. The table also shows the participation in a voluntary "literature-workshop" (LW) for each student.

Point	10	14	7	18	12	15	12	12	8	8	4	9	21	11	15
LW	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No

***A)* 1 P**
Set up a cross tabulation where the variable "passed" is combined with the variable "participation in the literature work shop".

Using the table set up under Problem 1, I end up with the table below.

Points	4	7	8	9	10	11	12	14	15	18	21	Total
Frequency	1	1	2	1	1	1	3	1	2	1	1	15
Participate LW		Yes			Yes		1 Yes	Yes	1 Yes	Yes		6

The cross table can be set up as:

	Not passed	Passed	Total
LW	1	5	6
No LW	4	5	9
Total	5	10	15

***B)* 1 P**
Set up a suitable percentage distribution of data and provide an interpretation of the result with regard to examine if a relation exists between participation in the literature workshop and passing the exam of the German language course.

There are several possibilities. I have done as follows:

	Not passed	Passed	Total
LW	16.7 %	83.3 %	100 %
No LW	44.4 %	55.6 %	100 %
Total	33.3 %	66.7 %	100 %

About 83 % of the students who participated in the literature workshop also passed the German language course. For students who did not participate in the literature workshop the rate of passing the German language course was to about 56 %.

So it seems as participation in the literature workshop increases the chances of passing the language course in German. In this case *dependence* is observed.

***C)* 1 P**
What values should have been present in the table if the table should showed that there were no relation between participation in the literature workshop and passing the language course in German?

This is the case of *independence.* A table could look as:

	Not passed	Passed	Total
LW	3	3	6
No LW	4	5	9
Total	7	8	15

The distributions between passing and not passing and the LW variable are more identical.

Nils Karl Sørensen ©
Statistics II, BA-INT

Set 6: Hypothesis Testing

by Nils Karl Sørensen

Outline *page*

1. The Concepts of Hypothesis Testing

Hypothesis testing is properly the most used procedure in statistics.

- We use hypothesis to provide evidence in favor of some given *statement.* This statement is called a hypothesis.
- Hypothesis testing may be based on evidence either with regard to the mean or with regard to the variance. We shall work with the latter issue in the notes on the analysis of variance
- Hypothesis testing can be undertaken in total populations as well as in samples. In the first case we say that the standard deviation σ is known whereas in the latter case it is said that σ is unknown. This this case σ is replaced with *s*. This notation is consistent with the natation in notes set 4 from the course Statistics I

By undertaking a hypothesis test, we examine if a specific claim is valid or not. We compare a characteristic of a data set with "something different". In hypothesis testing this is called "comparing the null (H_0) with the alternative (H_a or H_1)".

So let us define the hypotheses as:

- The ***null hypothesis*** denoted H_0. This is *status quo*.
- The ***alternative hypothesis*** denoted H_a or H_1. This is the statement that will be accepted only if there is statistically convincing evidence that it is true.

By the use of statistical procedures we may find the alternative so convincing that it is of ***significance.***

We conduct a statistical test by undertaking the following procedure:

1. Identification of the problem a setting up the hypothesis
2. Select the level of significance α (by default set $\alpha = 0.05$ i.e. 95 percent)
3. Choice of the relevant statistical test (tester)
4. Calculating the tester
5. Evaluation of the outcome of the tester
6. Acceptance of either the null (H_0) or the alternative (H_1) (by use of the p-value)

It can be stated that the *level of significance* expresses the degree of *difference* acceptable in a given situation. For example if a level of significance is 95 percent and the H_1 is accepted then the H_1 hypothesis is so different from the base scenario H_0 that this event only will occur in 5 percent of all cases considered.

This way of thinking is very similar to the examining of a *statement* outlined in set 4 of the notes to Statistics I. We compared a given statement with the confidence level. If the statement was outside the confidence interval it was rejected as being false.

This looks simple, but there are some methodological issues here to be addressed. Let us go to court and look at a trial as an example!

Illustration and Discussion of Hypotheses and Outcomes

A person crosses an intersection downtown Sønderborg with traffic light in a situation where light is red. As a point of departure he is not guilty[1]. This is the base scenario or H_0.

The person is brought to court and is not guilty! So in this situation *H_0 true* and we do *not reject H_0.* This is so because in this situation he did not do it.

This looks nice and easy. However, in the court the judge has to listen to the arguments put forward by the defense as well as by the prosecute counsel. Then based on the arguments a decision is made. If the person is guilty a conviction is made. So the outcome of the trial is not given in advance. It depends on the arguments.

The hypotheses can be stated as:

H_0: Not guilty in crossing the intersection for red light
H_1: Guilty in crossing the intersection for red light

[1] This is the default in a democracy. You are not guilty, until your guilt in a crime is proved ☺

However, it is also possible that the person *did* cross the traffic light when red light was present. In this case the person is guilty! So in this situation H_1 is true and we *do not reject* H_1.

These two cases are nice because in each situation we make a correct decision. Our findings can be summarized in the table below.

	State of nature	
	H_0 is true	H_0 is false
H_0 is accepted	Correct decision	Type II error (β)
H_0 is rejected	Type I error (α)	Correct decision

As observed, two more nasty situations may prevail! Let us first consider a situation where a person is found guilty although the person was not guilty. In this situation, *H_0 is true*, but *H_0 is rejected*. This is a situation of miscarriage of justice or judicial murder! This could happen if an eye-witness not was able to identify the victim correctly or simply hates the person! Committing this error is called a ***type I*** or an ***α-error***.

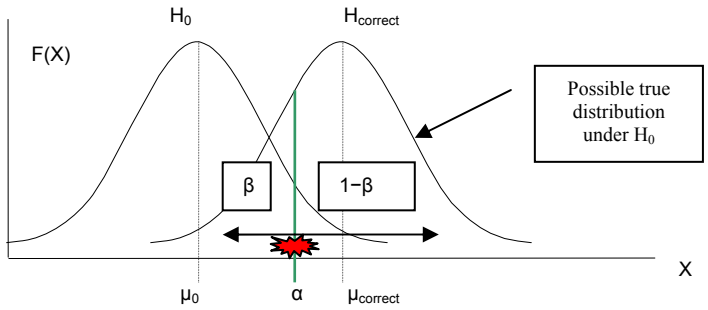
Is judicial murder the worst case? Well, let us consider another and perhaps more serious situation with a serial killer and the possibility of a type I or α-error. The completely innocent person has just been found guilty and is send to 10 times life time prison! However, the real serial killer is still around us and may be continuing killing!

Now it could be that we *actually* have found the real serial killer, and brought the person to court. However, evidence was weak, and the person was found not guilty, and could leave court as a free person, and continue killing (This is actually, what happens in the novel by Bret Easton Ellis *American Psycho)*! In this situation, *H_0 is false*, but *H_0 is accepted*.

This final situation is called a ***type II*** or ***β-error***. In the real world α-errors as well as β-errors are very serious, but in statistics the latter may be the worst. Why? It is because we are actually assuming that the wrong model or statement is the correct one.

For example in a game of ice hockey we may expect that the goal keeper takes 50 % of all shots on goal. Therefore, this is our model under H_0. Then if our team can make a goal in 55 % of all attempts on goal, we are likely to win the match.

However, it may turn out that the keeper from the other team took 60 % of our shots on goal, and consequently we are likely to lose the match. Our prior or model with regard to the goalkeeper of the other team was wrong! He was better than we expected. This error is serious from a statistical point of view, because the H_0 model is the wrong one, and our reference point (H_0) is therefore not consistent.



The illustration attempts to describe the situation. To the left we consider the model under H_0 where we expect the keeper to take μ_0 equal to 50 % of all shots on goal. If we are to the left of α, then there is no problem. We make goals in 55 % of our attempts and win the match. However, the true model is the one illustrated to the right with $\mu_{correct}$ equal to 60 %. Here we lose the match.

The *power function* is an attempt to express the possibility of *not committing a type II or β-error.*

Below let us finally summarize the findings from the traffic light and serial killer examples in a less statistical context.

	State of nature	
	Not guilty	Guilty
Not guilty	Correct decision	Type II error (β)
Guilty	Type I error (α) Judicial murder	Correct decision

The two types of errors α and β should be minimized. Observe that:

α = P(type I error) = P(reject H_0 | H_0 is correct)
β = P(type II error) = P(accept H_0 | H_0 is false)

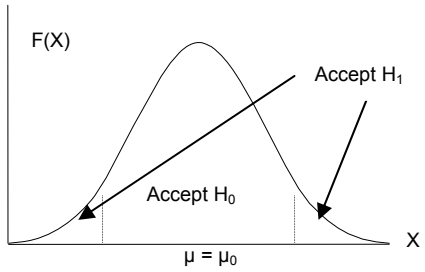
The *power function* (i.e. the strength of our model) is then defined as (1−β).

Formulation of Hypotheses

Let us consider an example of a test for the mean μ being equal to a specific value. Such a test may be set up in three ways.

I) Two-sided test H_0: $\mu = \mu_0$ (the mean is equal to the given value)
H_1: $\mu \neq \mu_0$ (the mean is *not* equal to the given value)

Illustration:

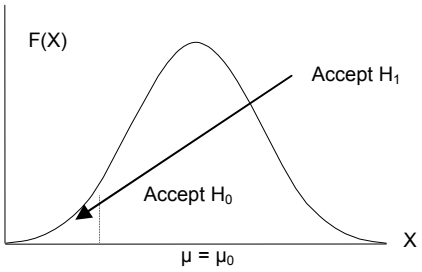


This situation is similar to the case that was considered in the set of notes on *confidence intervals* in the course Statistics I. If the statement is outside the confidence interval then H_0 is rejected and H_1 is accepted.

This case is in many cases not efficient. It case I the statement *different from* is examined. In many cases a higher degree of specification is required. This brings us to the next two cases.

II) Left-tailed test H_0: $\mu \geq \mu_0$ (the mean is equal to or larger than the given value)
H_1: $\mu < \mu_0$ (the mean is smaller than the given value)

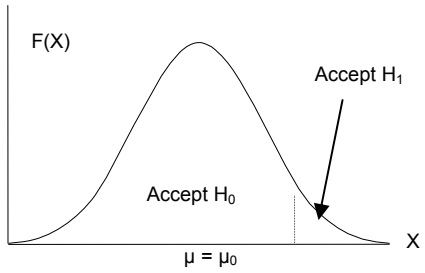
Illustration



How can the hypothesis be formulated correctly? As a rule of thumb, start with formulating the alternative hypothesis. This is in many cases the easiest thing to do.

III) Right-tailed test $H_0: \mu \leq \mu_0$ (the mean is equal to or smaller than the given value)
$H_1: \mu > \mu_0$ (the mean is larger than the given value)

Illustration



As evident and also stressed above hypothesis testing has a lot in common with setting up a confidence interval.

- In the *two-sided case* the probability of accepting H_0 is equal to $(1-\alpha)$ and the probability of accepting H_1 is equal to $\alpha/2$ (upper and lower tail)
- In the *one-sided case* the probability of accepting H_0 is equal to $(1-\alpha)$ and the probability of accepting H_1 is equal to α (upper or lower tail)

2. Z-Based and T-based Tests about a Population Mean

Z-based Test with σ known (Total Population Test)
Let us consider a case where we want to provide a test for the mean in a situation where the population standard deviation σ is known from the total population. So we are dealing with a total population data set or with a large data set. If we assume a *normally distributed data set* then we can define a Z-test.

Consider for the present purpose the two-sided case.

Define the hypotheses: $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Then the tester is equal to:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Z is the value from the normal distribution, σ is the standard deviation, n is the number of observations in the dataset, μ_0 is the mean under H_0. This is the mean from the dataset. Finally, X is the statement value to be examined. The formula above is an extension of the

formula for the confidence interval, and derived from the *central limit theorem.* To make things (a little) more clear consider the example in the next Section.

Introducing the P-value and an Example

Go back to the case with the price of gasoline considered in the notes to Statistics I on probability as well as the notes on confidence intervals.

Remember that our analysis was based on a normally distributed data set with the mean equal to 9.95 DKK, the standard deviation equal to 0.30 DKK and a data set with 25 observations.

- Consider the claim that the price of gasoline at a given day is equal to 10.10 DKK. How likely is this?

Let us state the hypotheses. As we have said noting specifically about the design of the test we assume that we consider a two-sided test and we conduct the test at a 95 % level of significance. This is the most commonly used assumptions. So

H_0: $\mu = 10.10$ (the price is equal to 10.10 DKK)
H_1: $\mu \neq 10.10$ (the price is *not* equal to 10.10 DKK)

The notation is as follows: $\overline{X} = 10.10$, $\mu_0 = 9.95$, $\sigma = 0.30$ and $n = 25$.

Consider the hypothesis in another way, and inspect the difference $D = \overline{X} - \mu_0$. Written in differences the hypothesis looks as

H_0: $D = \overline{X} - \mu_0 = 0$
H_1: $D \neq \overline{X} - \mu_0 \neq 0$

Because of the *central limit theorem* the following will be true under H_0

$$\overline{X} \approx N(9.95\,;\frac{\sigma^2}{25})$$

As $\overline{X}$ and μ both are normally distributed then D must also be normally distributed. With our parameters, and using the central limit theorem as well as the rules for discrete random variables, under H_0 it must be true that

$$D = (10.10 - 9.95) \approx N(0\,;\frac{0.30}{\sqrt{25}}) \qquad \text{(under } H_0\text{)}$$

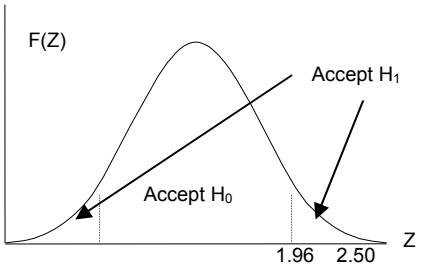
Using the expression for Z above then

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{10.10 - 9.95}{0.30 / \sqrt{25}} = \frac{0.15}{0.06} = 2.50$$

This value is then approximately $\approx N(0;1)$ under H_0.

Assuming that $\alpha = 0.05$ i.e. a 95 percent level of significance, and a two-sided test, then $\alpha/2 = 0.025$ and $Z = \pm 1.96$. We find the Z-value by use of **Statistics Tables.** In the notes on confidence intervals it was shown how the value 1.96 was obtained.

Illustration



It is found that $1.96 < 2.50$. The conclusion is that H_0 is rejected. The price of 10.10 DKK is significantly different than the price of 9.95 DKK. The price is then too high.

Generally, what is the interpretation of the value of Z, when Z = 2.50 as found in the example? This is the value where H_0 is rejected. Below Z = 2.50, H_0 is correct, and above 2.50, H_1 is correct. ***Therefore, it is the value where we will commit a type I error (α-error).*** This value is denoted the *p-value* or *prob-value*. This value is defined as

We can interpret the ***p-value*** as:

- The probability, computed assuming that the null hypothesis H_0 is true, of observing a value of the test statistic, which is at least as extreme as the value actually computed from the sample data. *(We reject H_0 in favor of H_1 at the level of significance α, if and only if the p-value is less than α).*

- The smaller the p-value, the less likely are the sample results, if the H_0 is true.

If the p-value for testing H_0 is *less* than

- $p < 0.10$ we have ***weak significance*** that H_0 is false
- $p < 0.05$ we have ***significance*** that H_0 is false
- $p < 0.01$ we have ***strong significance*** that H_0 is false

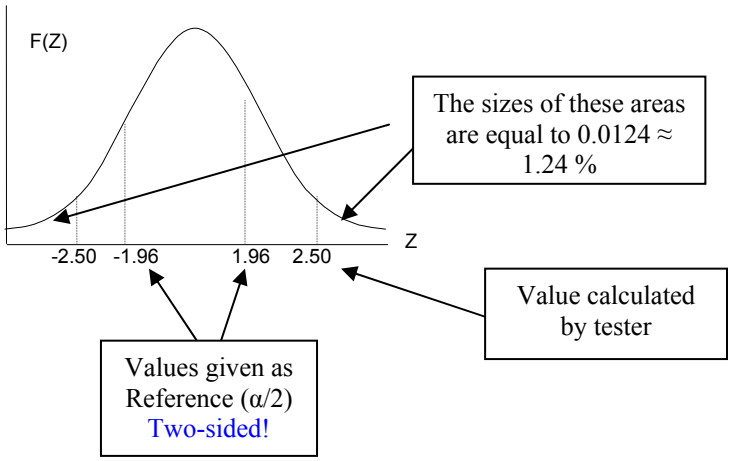
Let us as an example calculate the p-value in our example and use the **Statistics Tables** for the Normal distribution:

$$P(Z < -2.50) + P(Z > 2.50) = 2P(Z < -2.50) = 2(0.0062) = 0.0124 \ \ (=p)$$

We multiply by two because we have set up a two-sided test. If we had used a one-sided test the *p-value* would only have been half the size!

What does the p-value tells us? Look at the box on the previous page. We accept H_1 at the $0.01 < 0.0124 < 0.05$ (5 % level), so we have significance. We do however *not observe strong significance.* The value of 10.10 DKK is significantly different from 9.95 DKK, but it is not strongly different.

Illustration:



T-based Test with σ unknown (Sample Test)
Along similar lines we can set up a test in the case where the population standard deviation *not* is known. Since the total population normally not is known this will be the typical case especially in small samples. In this case we use a very similar tester, but now based on the ***t-distribution***.

Again consider a two-sided case.

Define the hypotheses: H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Then the tester is equal to:

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$$

The tester will be *t-distributed* with **degrees of freedom** equal to $df = (n-1)$.

What is the t-distribution and why should it be considered?
In most of the cases considered in statistics the applied dataset is small for example with 25 observations as in case with the price of gasoline. The *central limit theorem* tells us that the standard deviation in the sample is a transformation of standard deviation in the total population, and that the sample standard deviation is *smaller*. In the sample the uncertainty on the mean is consequently smaller.

If for example samples are taken from a production line in a brewery in order to control that the amount of centiliters filled on the bottles is correct, then the central limit theorem tells us that the uncertainty on the standard deviation is *smaller* than for the production in total. This will lead to a systematic higher degree of rejection of bottles with too much or too little content of beer than should be expected.

This problem was in 1908 considered by a scientist and statistician W.S. Gossett at the Guinness brewery in Dublin. To solve the problem he developed a transformed version of the normal distribution. The Guinness brewery, however, did not allow its employees to publish findings under their own name or in the name of the brewery. Therefore, Gossett published his findings under the pen name *Student-t.* As a result, the distribution became known also as the **Student-t distribution**.

The *t-distribution* is a transformed bell shaped normal distribution. The shape of the distribution changes as the number of observations increases. There is a t-distribution for each degree of freedom. We have lost one degree of freedom, because we used some information when we calculated the mean for the total population. See also on *degrees of freedom* in earlier notes to Statistics I.

As the t-distribution is a transformed normal distribution the mean of the distribution is zero (it is symmetric around zero). So the t-distribution acan take negative as well as positive values.

For degrees of freedom larger than 2, the ***variance*** of the t-distribution is equal to $\boldsymbol{df/(df-2)}$. Here *df* is equal to degrees of freedom.

The t-distribution is tabulated in **Statistics Tables** page 10. An extract of the table can be found on the next page. Notice that due to space only positive values are shown.

Observe that when the number of degrees of freedom increases the ***t-distribution approaches to a normal distribution.*** Setting degrees of freedom equal to infinity (or ∞) this is evident *from the bottom line in table of critical values.*

	t Values				
df	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
…	…	…	…	…	…
…	…	…	…	…	…
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
…			…	…	…
…	…	…	…	…	…
∞	1.282	1.645	**1.96**	2.326	2.576

How is the table read? In the case with the gasoline there were 25 observations. The number of degrees of freedom is then equal to *dg = (n−1) = 25−1=24.* For a two-sided test the t-value is equal 2.064. This is at α/2 = 0.05/2 = 0.025.

What if it was a one-sided test? Then α = 0.05. From the table the t-value is then found to be equal to 1.711.

A very frequently asked question is: "When is a sample large?" I am not to judge! In the old days the text books said 30 observations. With this number of observations mean and standard deviation could be computed manually within a traceable amount of time. With the theoretical distributions observed from the tables and with our easy access to computer power today a value equal to 120 observations is more likely.

The Example above reconsidered, but now with t-distribution

Look now at the ***t-test***. Then it is assume that we are dealing with a sample of an endless process of visits to the gas station (sounds fair)! Then the tester is

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{10.10 - 9.95}{0.30/\sqrt{25}} = \frac{0.15}{0.06} = 2.50 \qquad \text{(as above)!}$$

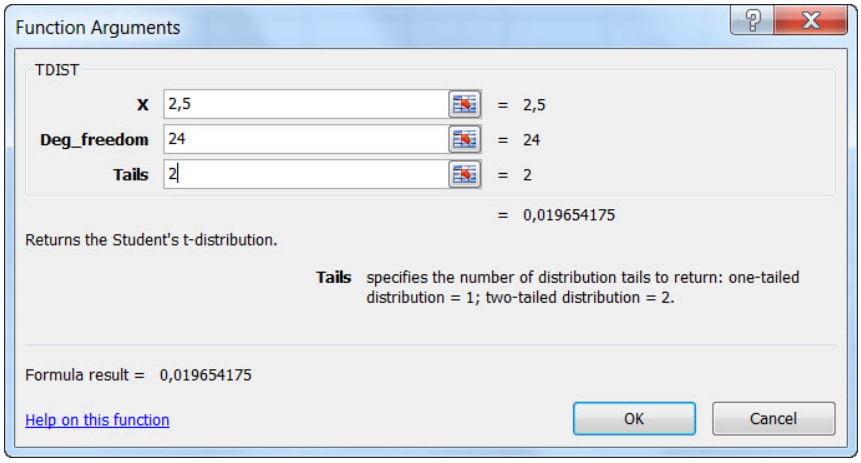
This tester is then *t-distributed* with *degrees of freedom* equal to *df = (n−1) = (25−1) = 24.* Assuming that α = 0.05, and a two-sided test, then α/2 = 0.025 we find the t-value by use of the table of the *t-distribution* in the **Statistics Tables**. Here *t = 2.064* as just explained above.

As *2.064 < 2.50* we still have the same conclusion, namely that H_0 is rejected and H_1 is accepted.

What is the *p-value* in this case? This is not so easy to answer as above. This is so because we do not have the full distribution of the t-statistic for each degree of freedom.

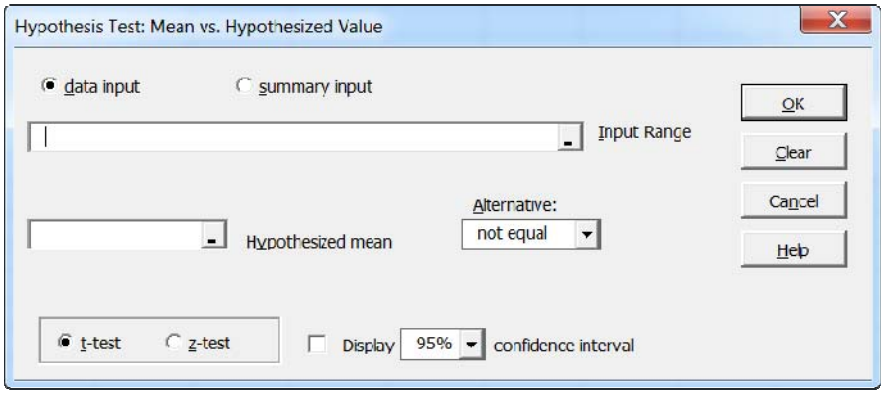
From the table above it can be observed that at α=0.01 a value equal to 2.492 is observed at 24 degrees of freedom. This is very close to 2.50. As 2.492<2.50 the p-value is just below 0.01, so strong significance is found.

We can find the *exact p-value* of the t-test in two ways. Either by interpolation or by use of Excel. In Excel type **fomulas | insert function | statistical | tdist** and insert 2.5 for X. Remember that 2 = two-tailed and 1 = one-tailed. Then find *p= 0.019* as evident from the screenshot below



Megastat and simple hypothesis testing
The tests for mean outlined in this section can also be conducted by use of Megastat.

Open Megastat and select *Hypothesis tests* then select *Mean vs. hypothesized value*. The following dialog box will appear:



Load the relevant dataset, and state the value of the hypothesized mean. Under *alternative* the type of the hypothesis can be stated i.e. not equal, greater than (right sided) or less than (left sided). In addition it can be stated if a z-test or a t-test should be performed (t-test is the default).

The pocket calculator and simple hypothesis testing
The tests presented in this section can also be performed by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → "1: Z-Test" → ENTER, and insert the values for μ_0, $\bar{X}$, σ and n. Remember to mark the type of hypothesis to be considered. Finally use CALCULATE → ENTER. The results from above the result.

In case of a t-test use instead Tast STAT → TESTS → "2: T-Test" → ENTER. As before, but now s_x=0.30.

3. Z-based Tests about a Population Proportion

In the case with a large sample we can set up a test for the population proportion along similar lines as above. This is always a large sample test, so the number of observations is large. As a result this is always a Z-test.

This test is just an extension of the confidence interval for p defined in the notes on confidence intervals in Statistics I.

Consider the hypothesis: $H_0: p = p_0$
$H_1: p \neq p_0$

Define then the tester

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Here $\hat{p}$ is the statement, p_0 is the share under H_0 and n is the number of observations.

Example: The backer from Langeland
Let us go back to the baker from the island of Langeland that we worked with in the notes on confidence intervals in Statistics I.
Let us claim that the local mayor of one of the three municipalities predicts that 63 % of the votes will vote "yes" for the union of the municipalities. Remember that 385 cookies out of a total of 656 cookies were decorated with a "yes". The share under the null is then equal to $p_0 = x/n = 385/656 = 0.587$.

- Use the statistics based on the baker sales of cookies, and set up a test at 5 % level of significance. Examine the claim put forward by the mayor.

Initially set up the hypothesis: H_0: $p = 0.63$ (claim by mayor is correct)
H_1: $p \neq 0.63$ (claim by mayor is false)

Define then the tester

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.63 - 0.587}{\sqrt{\frac{0.587(1-0.587)}{656}}} = \frac{0.043}{0.0192} = 2.24$$

As $\alpha = 0.05$ and $(1-\alpha) = 0.95$. If a two-sided test is conducted, then $\alpha/2 = 0.025$, and we find the value of Z to be equal to ± 1.96.

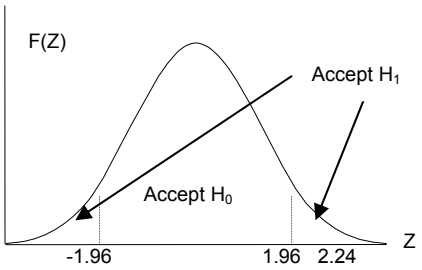
As $1.96 < 2.24$ we reject H_0. Then 0.63 is significantly different from 0.587, and the *statement put forward by the mayor is wrong.*

Now what is the *p-value?* We find that by use of the table for the normal distribution in **Statistics Tables**. Then:

$$P(Z < -2.24) + P(Z > 2.24) = 2P(Z < -2.24) = 2(0.0125) = 0.0250 \ \textit{(=p)}$$

So $p < 0.05$ and H_1 *is accepted.*

Illustration:



However, if a 1 % level of significance had been used then $p > 0.01$ and the conclusion would have been reverse. The outcome of the test is sensible to the level of significance.

Notice, also that the result of the test at the 5 % level could have been obtained by inspection of the 95 % confidence interval calculated in the notes on confidence intervals from Statistics I. Here it was found that $[0.549\,;0.625]$. As the value 0.63 falls outside the range of the interval the statement must be false.

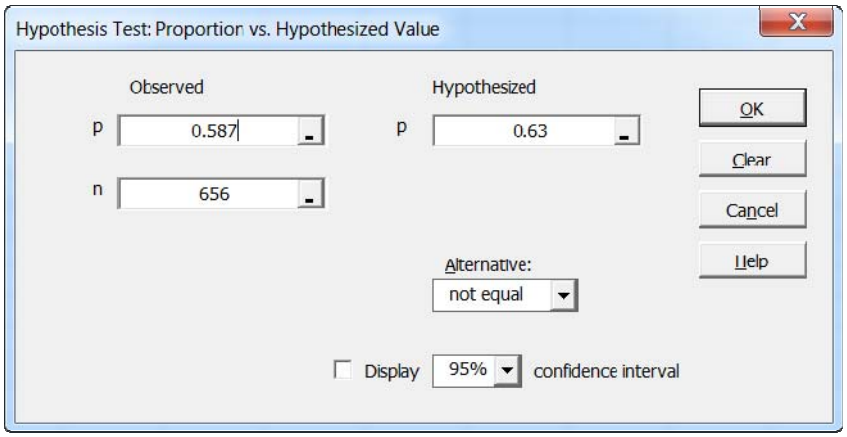
Let us finally consider this case as a one-sided test. Then $\alpha = 0.05$, and the associated Z-value is equal to 1.645. As observed this will not change the conclusion as $Z = 2.24$.

Do we always obtain a similar outcome? No, consider the where $1.645 < Z < 1.96$. In this case we accept H_1 in the one-sided case, but accept H_0 in the two-sided test. This underlines the importance of formulating the test correct.

Conducting the test by use of Megastat

The tests for population proportion can also be conducted by use of Megastat.

Open Megastat and select *Hypothesis tests* then select *Proportion vs. hypothesized value*. The following dialog box will appear:



The values given have been inserted. The following output will result:

Hypothesis test for proportion vs hypothesized value

Observed	*Hypothesized*	
0.587	0.63	p (as decimal)
385/656	413/656	p (as fraction)
385.072	413.28	X
656	656	n
	0.0189	std. error
	-2.28	z
	.0225	p-value (two-tailed)

The result is a little different. This is due to the set up. It is not relevant relative to the exam ☺

The pocket calculator and testing for population proportion

The test presented in this section can also be performed by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → ”5: 1-PropZtest” → ENTER and insert the values for x=385, n=656 and $\hat{p}$=0.63. Finally use CALCULATE → ENTER. A result similar to the one obtained by Megastat will appear[2].

[2] On the screen of the calculator it says ”p_0”. This is an notation error.

Nils Karl Sørensen ©
Statistics II, BA-INT

Set 7: Statistical Inference Based on Two Samples

by Nils Karl Sørensen

Outline

page

Sections marked with an * will not be subject to independent exam questions.

1. Comparing Two Population Means

We extend the analysis to include two data sets. We assume that we consider two independent random stochastic processes labeled by subscript *1* and *2* respectively. Further, we assume that the two data sets are normally distributed. The two populations or samples are *at most* of different size with observations n_1 and n_2 respectively, so $n_1 \neq n_2$.

Comparisons of two data sets for equal mean, proportion share or variance is very frequently used in marketing, production control or medicine. For example

- We examine the impact of a market campaign by comparison of the markets shares before and after the campaign
- We examine mean of a production process to test for product homogeneity
- We examine the impact of a new production process
- We examine the impact of a new type of medicine by comparison of a group who have received the new type of medicine by a reference group
- We examine if the volatility of two types of for example stocks is similar. The stock with the lowest volatility is has the lowest risk

As the case in the previous set of notes, we again consider two cases. First, we deal with the case where σ is known, and thereafter we consider the case where σ is unknown. Finally, we consider a test for equal variances. This test is a little special because two squared variables is considered σ^2.

Z-based test with σ known

In this case, we consider a large sample data set or a total population. So the normal distribution is valid. Again we look at the two-sided case. In the exercises we will also work with some one-sided cases.

Hypotheses:

H_0: $\mu_1 = \mu_2$ (The mean in the two sets are equal)
H_1: $\mu_1 \neq \mu_2$ (The mean in the two sets are different)

or

H_0: $\mu_1 - \mu_2 = D_0$ (The difference among the means is equal to D_0)
H_1: $\mu_1 - \mu_2 \neq D_0$ (The difference among the means is different from D_0)

Where D_0 is some hypothesized difference between the two data sets. Normally, this takes the value of zero.

How can we use D_0? D_0 states that under H_0 there is a difference in level between the two data sets. This could for example be wage by gender. Although many Western European countries claims that the wage rate among females and males are similar for similar types of labor the reality may be different. A survey may for example indicate that a gab by gender equal to 1,000 € per year exists on the average income. In such a case the hypotheses may be stated as:

H_0: $\mu_{male} - \mu_{female} = 1{,}000$ (the difference is equal to 1,000 €)
H_1: $\mu_{male} - \mu_{female} \neq 1{,}000$ (the difference is different from 1,000 €)

What is the interpretation? Under the null we inspect the stochastic variation in the average income *given* a difference equal to 1,000 €. Under the alternative we postulate that the average income is different *although* a difference equal to 1,000 € is included. Accepting the alternative hypothesis indicates that the difference in the income by gender is either smaller or larger than 1,000 €. Notice that the hypotheses can be made more specific by setting up a one-sided test:

H_0: $\mu_{male} - \mu_{female} \leq 1{,}000$ (the difference is 1,000 € or less)
H_1: $\mu_{male} - \mu_{female} > 1{,}000$ (the difference is larger than 1,000 €)

Here a right-sided test is applied.

The Z-test is a straight forward extension of the test provided in the notes set 6, page 6 bottom. So

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

It is evident that

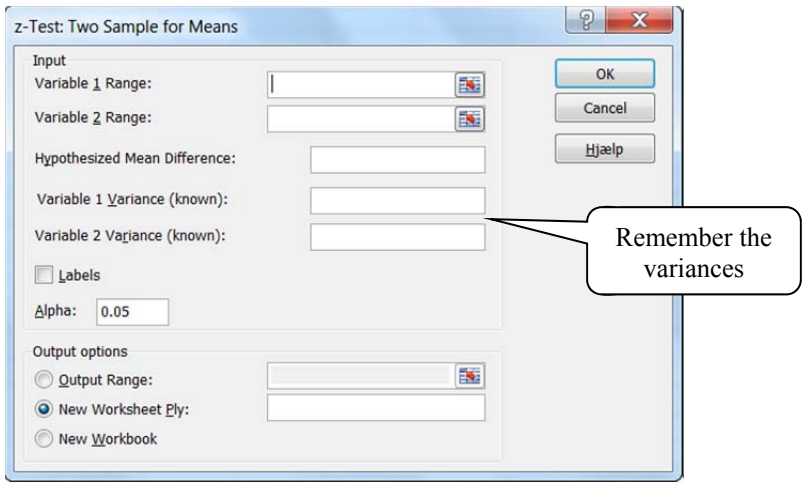
The mean is: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

The standard deviation is: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

A *100(1−α) percent confidence interval* for $\mu_1 - \mu_2$ is:

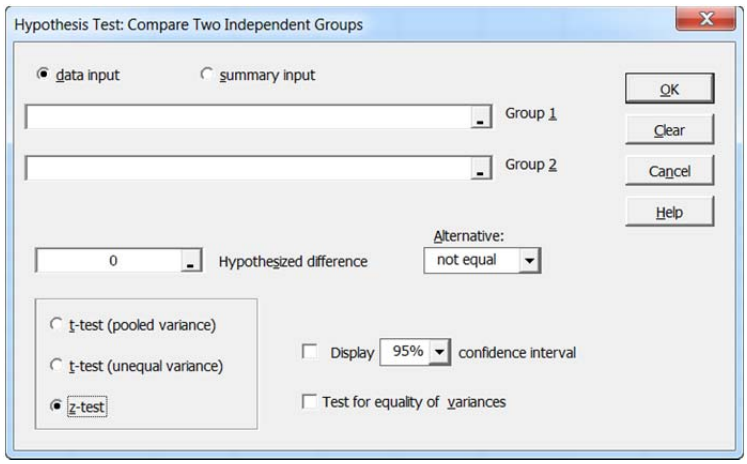
$$\left[(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

In **Excel** click **Data | Data analysis | Z-test: Two Sample for Means**. Click on "OK" and obtain the box below. This box needs a comment. As the standard deviations is assumed to be known you have to calculate them *before* application of the test. This task can be undertaken by use of the ***descriptive statistics feature*** in Excel. The interpretation of the output will be discussed later in these notes.



Megastat and Z-based hypothesis testing for two independent groups
The test for two equal means outlined in this section can also be conducted by use of Megastat.

Open Megastat and select *Hypothesis tests* then select: *Compare two independent groups*. The following dialog box will appear:



Load the relevant datasets (groups), and state the value of the hypothesized mean. Under *alternative* the type of the hypothesis can be stated i.e. not equal, greater than (right sided) or less than (left sided). In addition it can be stated if a z-test or a t-test should be performed (t-test is the default).

There are many options available. The t-test can be performed either with a pooled variance or with unequal variance. The former is the normal, but both types of test will be considered below. It is also possible to insert a *hypothesized difference* like for example the 1,000 € in the example above.

Finally, it is possible to perform a *test for equality of variances* among the two datasets. This test will be considered in Section 4 in this set of notes.

The pocket calculator and Z-based test among two independent groups
The tests presented in this section can also be performed by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → ”3: 2-SampZTest” → ENTER, and insert the values for $\bar{X}_1, \bar{X}_2$, σ_1, σ_2, n_1 and n_2. Remember to mark the type of hypothesis to be considered. Finally use CALCULATE → ENTER.

T-based test with σ unknown
Along similar lines we can set up a test in the case where the population standard deviation is *not* known. This will normally be the case as we seldom know the exact size and shape of the total population.

Let us again consider the two-sided case. As the standard deviations not are known we use the ***t-distribution*** for the test. This distribution was introduced in note set 6 on hypothesis testing.

A special problem will be present for the t-test namely the treatment of the sample variances for the two dataset. The variances can either be assumed to be identical or they can be assumed to be different. We shall consider each case in turn. The outcomes of the two tests are frequently similar. The test assuming equal variances is the most straightforward to perform.

Case I: T-based test assuming identical variances among the two dataset
The hypotheses are as for the Z-test. The tester can be written as

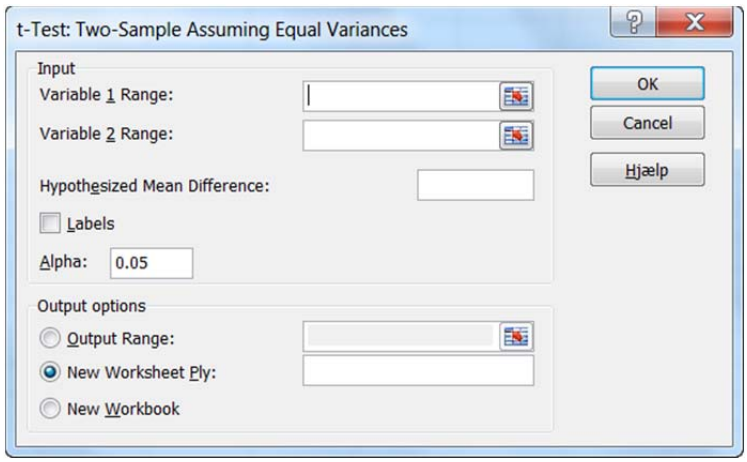
$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The tester will be t-distributed with degrees of freedom equal to $df = (n_1 + n_2 - 2)$. We subtract 2, because we have lost two degrees of freedom when we calculated the population mean in each of the two samples.

s_p^2 is the "pooled sample variance". This is a weighted average of the two sample variances. It is written as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

In **Excel** we find this tester by clicking **Data | data analysis | t-test: Two samples with equal variance.** The box looks as:



In this case it is easier than with the Z-test. We only have to specify the two data variables. No variances are needed!

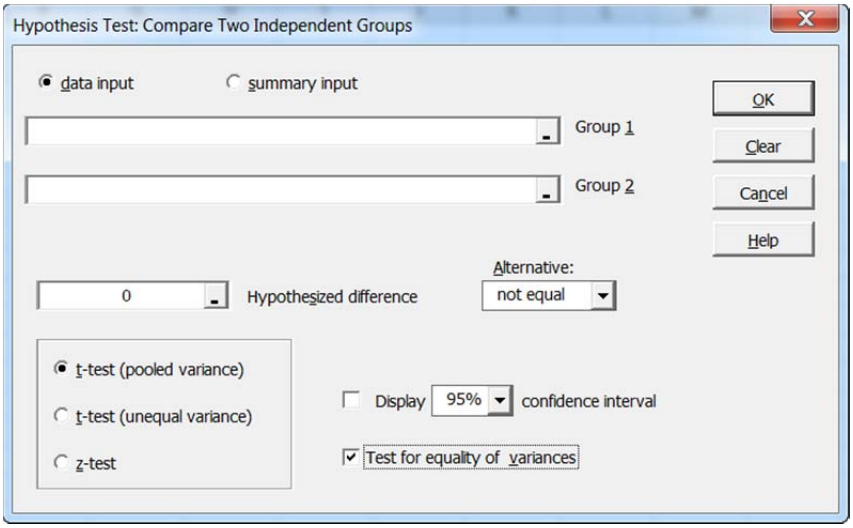
A *100(1−α) percent confidence interval* for $(\mu_1 - \mu_2)$ is

$$\left[(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$$

Degrees of freedom and the definition of s_p^2 are as above.

Megastat and t-based hypothesis testing for two independent groups
The test for two equal means outlined in this section can also be conducted by use of Megastat. This is very similar to the procedure outlined for the Z-test.

Open Megastat and select *Hypothesis tests* then select: *Compare two independent groups*. The following dialog box will appear:



As the case with the Z-test load the relevant datasets (groups), and state the value of the hypothesized mean. Under *alternative* the type of the hypothesis can be stated i.e. not equal, greater than (right sided) or less than (left sided). Mark for a *t-test (pooled variance)*.

In addition, the assumption of equal variances has to be tested. Therefore, mark *test for equality of variances.* This test will be considered in Section 4 in this set of notes.

A worked example of the tests can be found in Section 5 of these notes.

The pocket calculator and t-based test among two independent groups
The tests presented in this section can also be performed by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → ”4: 2-SampTTest” → ENTER, and insert the values for X_1, X_2, s_1, s_2, n_1 and n_2. Remember to mark the type of hypothesis to be considered. Mark also: Pooled: Yes. Finally use CALCULATE → ENTER.

Case II: T-based test assuming identical variances among the two dataset*
In this case the tester is very similar to the Z-test statistic described above. The tester is given as

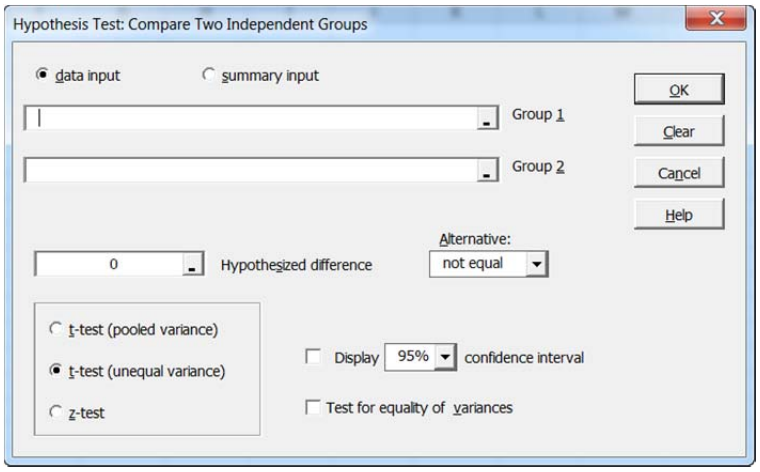
$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Now, the great problem is to find the number of degrees of freedom! You can say that the problem embodied in the formula above for the “pooled variance” is transformed to the expression for the degrees of freedom. Degrees of freedom is equal to

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2 / n_2\right)^2}{n_2 - 1}}$$

If *df* is not a whole number, we can round *df* down to the next small number! Wow – what an expression!

To perform this test ffor example open **Megastat** and select *Hypothesis tests* then select: *Compare two independent groups*. The following dialog box will appear:



Everything is a before. The only difference is that *t-test unequal variances* is marked. Notice that we do not have to perform a test for equal variances here.

In **Excel** we find this tester by clicking **Data | data analysis | t-test: Two samples with equal variance.** The resulting box is similar on the one displayed for Excel above.

On the **TI-84 pocket calculator.** Tast STAT → TESTS → "4: 2-SampTTest" → ENTER, and insert the values for X_1, X_2, s_1, s_2, n_1 and n_2. Remember to mark the type of hypothesis to be considered. Mark also: Pooled: No. Finally use CALCULATE → ENTER.

Finally, *a 100(1−α) percent confidence interval* for $(\mu_1 - \mu_2)$ is

$$\left[(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Degrees of freedom as defined above.

2. Comparing Two Population Proportions

We can extend the test conducted in Section 1 of these notes to be valid also for the comparison of two population proportions. When the sample sizes are large enough the distributions of the sample proportions $\hat{p}_1$ and $\hat{p}_2$ are both approximated well by a normal distribution. Then the difference between the two sample proportions will also be approximately normally distributed.

We can then set up the hypothesis:

H_0: $\hat{p}_1 - \hat{p}_2 = 0$	(the two population proportions are equal)
H_1: $\hat{p}_1 - \hat{p}_2 \neq 0$	(the two population proportions are different)

Or we can let D_0 substitute for zero, and be some hypothesized value of difference between $\hat{p}_1$ and $\hat{p}_2$.

As we are working with two stochastic variables there will also be two variances. This problem can be solved in various ways. First let us define the data sets. We have two samples of different size, so $n_1 \neq n_2$. The proportion shares are then $\hat{p}_1 = x_1 / n_1$ and $\hat{p}_2 = x_2 / n_2$ where $\hat{p}$ are the proportion shares, x is the observed number of outcomes with for example "yes", and n is the number of observations in the sample. The subscript referees to the sample number.

We can now calculate the "pooled share" as

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

A large sample tester can then be stated as

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where D_0 is hypothesized difference between the two proportions under H_0. Typically this difference is equal to zero. Alternatively the tester can be written as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sigma_{\hat{p}_1 - \hat{p}_2}}$$

What is different is the treatment of the standard deviation. If we use the expression for the standard deviation of a binominal distributed random variable, and remember that the standard deviation (or the variance) of difference among two independent random distributed variables is equal to the *sum* of the standard deviations, then we obtain

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

For the formulas above to be valid n_1 and n_2 are large enough if, n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are all at least 5.

We can use the formula just given to set up a *100(1−α) percent confidence interval* for (p_1-p_2)

$$\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-p_2)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

Example: The Backer from Langeland
In the previous set of notes we have worked with the backer from the island of Langeland and his cookies ☺

In this case consider the outcome of the election on February 12th 2003 regarding the unification of the 3 municipalities into a single municipality for Langeland. Out of the 3,400 recorded voters in the 3 municipalities 2,695 gave their vote. In all municipalities there were a majority of "yes" votes, so the union among the municipalities was decided. Across the municipalities the votes for "yes" equaled 1,665.

- Decide on a 5 % level how good the forecast provided backer actually was in predicting the outcome of the vote

Remember that the backer totally sold 385 "yes"-cookies out of 656. This equaled 58.7 %.

We denote the results by the backer with subscript 1, and the actual outcome of the vote with subscript 2. Then the shares of "yes" is equal to

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{385}{656} = 0.587 \qquad \text{and} \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{1665}{2695} = 0.618$$

The hypothesis to be examined is

H_0: $\hat{p}_1 - \hat{p}_2 = 0$ (the forecast was correct)
H_1: $\hat{p}_1 - \hat{p}_2 \neq 0$ (the forecast was *not* correct)

In the present case we consider a two-sided test with α=0.05.

First we calculate the "pooled share" (this has to the used in order to calculate the "pooled variance"):

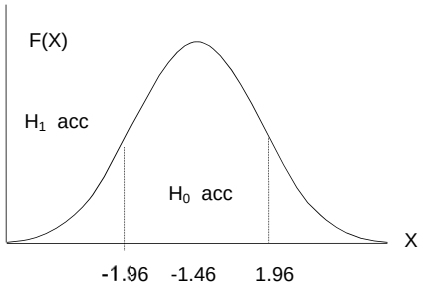
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{385 + 1665}{656 + 2695} = \frac{2050}{3351} = 0.612$$

We then calculate the Z-tester (with this sample size there is no problem with the normal distribution):

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.587 - 0.618}{\sqrt{0.612(1-0.612)\left(\frac{1}{656} + \frac{1}{2695}\right)}} = \frac{-0.031}{0.0212} = -1.46$$

The critical value for $Z_{\alpha/2} = Z_{0.025} = -1.96$. This is found in the table for Z in **Statistics Tables**. As $-1.96 < -1.46$ we accept H_0. So the forecast provided by the backer was a good predictor for the outcome of the vote!

Illustration



The assumptions for the test can also be examined:

$$n_1p_1 = 656(0.587) = 385 \qquad n_1(1-p_1) = 656(1-0.587) = 271$$
$$n_2p_2 = 2695(0.618) = 1665 \qquad n_2(1-p_2) = 2695(1-0.618) = 1029$$

As all values exceed 5 the assumptions are fulfilled.

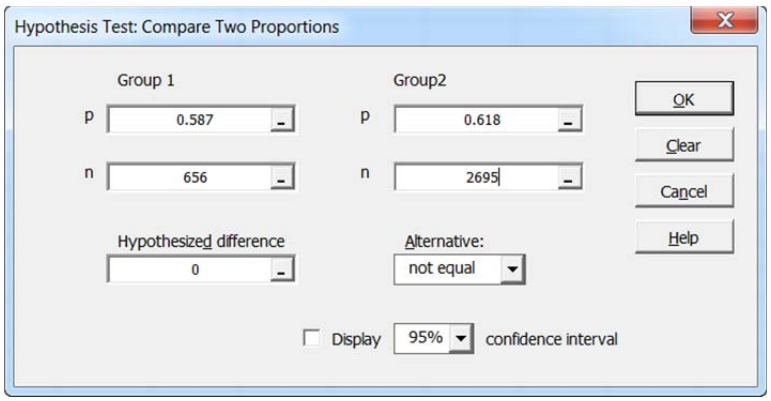
The *p-value* can be found as:

$$P(Z < -1.46) + P(Z > 1.46) = 2P(Z < -1.46) = 2(0.0721) = 0.1424$$

So this value is not even weakly significant.

Test for population proportion on Megastat and the pocket calculator

Open Megastat and select *Hypothesis tests* then select: *Compare two independent proportions*. The following dialog box will appear:



I have filled out with the data here. Click on OK and get the output:

Hypothesis test for two independent proportions

p1	*p2*	p_c	
0.587	0.618	0.6119	p (as decimal)
385/656	1666/2695	2051/3351	p (as fraction)
385.072	1665.51	2050.582	X
656	2695	3351	n
		-0.031	difference
		0.	hypothesized difference
		0.0212	std. error
		-1.46	z
		.1440	p-value (two-tailed)

The output confirms the results found above.

The tests presented in this section can also be performed by use of the **TI-84 pocket calculator.** Tast STAT → TESTS → ”6: 2-PropZTest” → ENTER, and insert the values for x_1, x_2, n_1 and n_2. Remember to mark the type of hypothesis to be considered. Finally use CALCULATE → ENTER.

3. Chi-based Confidence Intervals for the Population Variance

Back in Section 1 of these notes we considered a t-test for equal means among two independent dataset assuming that the variances of the two dataset were equal. This assumption has to be examined by a test. Looking at the box from Megastat on page 6 it is also evident that such a test can be performed because it is possible to mark for *a test for equality of variance.*

In order to develop a more formal procedure for a test assume two dataset labelled 1 and 2 respectively. The variances can be written as σ_1^2 and σ_2^2. Hypothesis for a test for equal variances can then be written as:

H_0: $\sigma_1^2 = \sigma_2^2$ (the 2 dataset has identical variances)
H_1: $\sigma_1^2 > \sigma_2^2$ (the variance in dataset 1 is the largest)

Under H_0 the variances are equal or *homogeneous.*

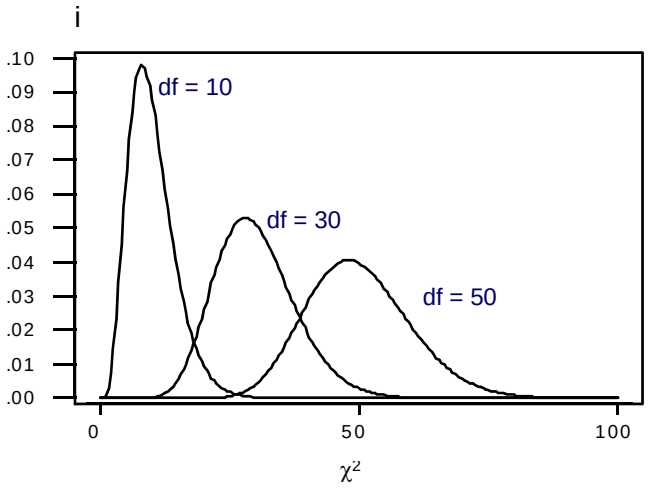
A special feature of the variance is that it is a square. The variance as well as the standard deviation will always be a positive number. This is a fundamental different situation that with the Z- or t-based tests. Here the tester can take a negative value. In order to deal with the special feature a set of new distributions has to be introduced.

A confidence interval for the variance and the standard deviation
Before moving to the test for equal variances or variance homogeneity a more simple case is considered namely to set up a confidence interval for the variance or standard deviation of a single dataset.

A confidence interval for the variance may be of interest when we want to investigate the volatility in for example production processes, inequalities among regions over time (this is called sigma convergence) or in finance in order to identify risky assets.

To set up a confidence interval for the sample variance or the standard deviation a *squared* statistical distribution is needed.

Chi-Square Distributions with 10, 30 and 50 Degrees of Freedom



The ***chi-square distribution*** labeled by χ^2 has this feature. Because it is a square, it cannot be negative. The probability curve of the χ^2 distribution is skewed to the right as shown in the illustration. As the number of degrees of freedom increases the distribution becomes more symmetric as displayed below and will approaches a normal distribution. The chi-squared distribution is also a transformed and squared Normal distribution, so this feature is expected.

The mean of a chi-squared distribution is equal to the degrees of freedom parameter *df*. The variance of a chi-squared distribution is equal to twice the number of degrees of freedom. If *df* increases the chi-square distribution approaches a normal distribution with mean *df* and variance $2\times(df)$.

We apply the chi-square distribution to problems of estimation of the population variance using the property saying that when sampling from a normal distribution the random variable

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom. The chi-square distribution links the ration among the sample variance s^2 and the total population variance σ^2.

The formula for the chi-squared distribution is tabulated in **Statistics Tables** page 11. As observed from the illustration on the previous page, especially for small samples the distribution is not symmetric. Therefore, we cannot use equal values with opposite signs (such as ± 1.96 as we did with the *Z*-value of the Normal Distribution) and instead we must construct the confidence interval using the two distinct tails of the distribution.

A (1−α) 100% confidence interval for the population variance σ^2 (where the population is assumed normal) is

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

where $\chi^2_{\alpha/2}$ is the value of the chi-square distribution with $n-1$ degrees of freedom that cuts off an area of $\alpha/2$ to its right and $\chi^2_{1-\alpha/2}$ is the value of the distribution that cuts off an area of $\alpha/2$ to its left (equivalently, an area of $1-\alpha/2$ to its right). In the case of the standard deviation, we just take the root. Let us work through an example!

Example: Confidence Interval on the Standard Deviation for the Price of Gasoline
Set up a 95 % confidence interval for the standard deviation of the price of gasoline that we worked with earlier for example in note set 6.

We have to calculate a 95 % confidence interval for the standard deviation. Remember that $n = 25$ and $s^2 = 0.30$. We use the formula above and the "chi-squared" distribution χ^2:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}}\right] = \left[\sqrt{\frac{(25-1)*(0.30)^2}{39.3641}}; \sqrt{\frac{(25-1)*(0.30)^2}{12.4012}}\right] = [0.2342\ ;\ 0.4173]$$

We set the number of degrees of freedom equal to $df = n-1 = 25-1 = 24$

From **Statistics Tables** we find the two χ^2-values to equal:

$$\chi^2_{0.025;24} = 39.3641 \text{ and } \chi^2_{0.975;24} = 12.4012$$

By substitution of these values, the confidence interval is obtained.

Let us look on how we found these values. An extract of the table in **Statistics Tables** page 11 will look as:

	χ values							
df	**0,99**	**0,975**	**0,95**	**0,90**	**0,10**	**0,05**	**0,025**	**0,01**
1	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349
2	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103
…	…	…	…	…	…	…	…	…
24	10,8564	12,4012	13,8484	15,6587	33,1962	36,4150	39,3641	42,9798
25	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	44,3141
26	12,1981	13,8439	15,3792	17,2919	35,5632	38,8851	41,9232	45,6417

A rule of memory for the chi-squared values for the confidence interval: The largest value namely the one to the right has to be used to the *left* in the formula for the confidence interval.

4. The F-distribution and Comparing Two Population Variances

With the knowledge on how to set up a confidence interval for the variance we are now in a position to undertake a test for comparing the variances among two datasets.

Use the hypothesis formulated in the previous section as a point of departure. Dividing the hypothesis by σ_2^2, it can be observed that the test is equivalent to:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{versus} \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$$

So if s_1^2 / s_2^2 is significantly larger than 1, we will reject H_0. Here s_1^2 is the variance of a random sample of n_1 observations drawn from a total population with variance σ_1^2, and s_2^2 is the variance of a random sample of n_2 observations drawn from a total population with variance σ_2^2.

What is the distribution describing the behavior of this ratio? The relation among σ^2 and s^2 was described by the chi-squared distribution for a single dataset. Here this distribution is applied, but now two datasets are present, so we are considering the ration among two independent datasets.

The ratio among two chi-squared distributed variables is described by the ***F-distribution*** named after the English statistician Sir Ronald A. Fisher.

The chi-squared distribution was similar with the t-distribution with respect to having a distribution for each degree of freedom. As we in the F-distribution are working with the ratio of two variables that may have different degrees of freedom the resulting distribution must be in two dimensions with regard to degrees of freedom.

Let us denote the degrees of freedom by k_1 and k_2. An F-distributed random variable is then

$$F_{(k_1,k_2)} = \frac{\chi_1^2 / k_1}{\chi_2^2 / k_2}$$

In Section 3 of the present notes, the chi-squared distribution was defined as

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

By substitution into the expression for the F-distribution

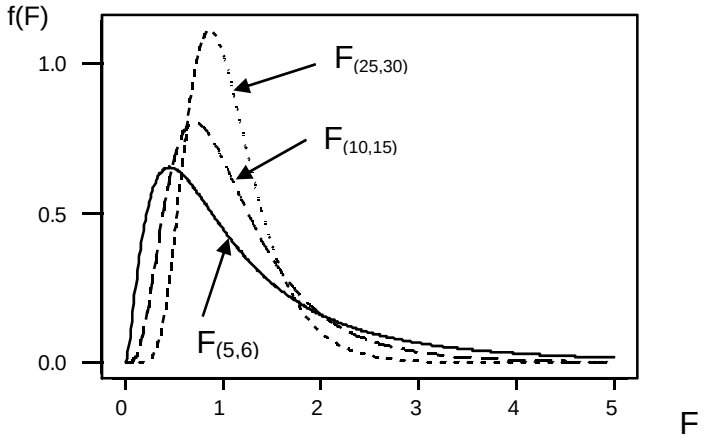
$$\frac{s_1^2}{s_2^2} = \frac{\chi_1^2\sigma_1^2/(n_1-1)}{\chi_2^2\sigma_2^2/(n_2-1)} = F_{(n_1-1,\,n_2-1)} = F_{df_1,df_2}$$

This useful distribution is tabulated in **Statistics Tables** pages 12−14. With two dimensions, the tables are rather space consuming, and rather complicated to read!

As the F-distribution is a transformed chi-distribution, it has similar properties with regard to the shape. This is evident from the illustration at the end of this Section. As the number of degrees of freedom increases the F-distribution becomes less skewed to the right, and the distribution will approach a normal distribution. As both df_1 and df_2 approaches to infinity the F-distribution takes a value equal to 1 regardless of the value of α (this is logically, because we are dealing with a ratio of two numbers (variances)).

Examples of F-distributions are given below. Observe that as the number of degrees of freedom increases the distribution becomes more symmetric just as the case with the chi-squared distribution and the Normal distribution.

F Distributions with different Degrees of Freedom



To sum up consider an extract from page 13 in **Statistics Tables.**

Table of F-Values 0,025

This Table was generated by use of the Excel function FINV

df 2 ↓ \ df 1 →	1	…	9	10	15	…	∞
1	648	…	963	969	985	…	1018
2	38.51	…	39.39	39.40	39.43	…	39.50
3	…	..	…	…	…	…	…
9	7.21	5.71	4.03	3.96	3.77	…	3.33
10	6.94	5.46	3.78	3.72	3.52	…	3.08
15	6.20	4.77	3.12	3.06	2.86	…	2.40
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
∞	5.03	3.69	2.12	2.05	1.84	1.21	1.00

The value of α is equal to 0.025, so the upper of a two-sided 95 % test is considered. The degrees of freedom are called df_1 and df_2. In this case df_1=10 and df_2=15. The associated value of the F-distribution is equal to 3.06.

For example assume a two datasets. The first data has 11 observations and the variance is equal to 3.1. In the second dataset has 16 observations and the variance is equal to 1.7. Are the variances among the two datasets equal?

Under H_0 the F-tester will be

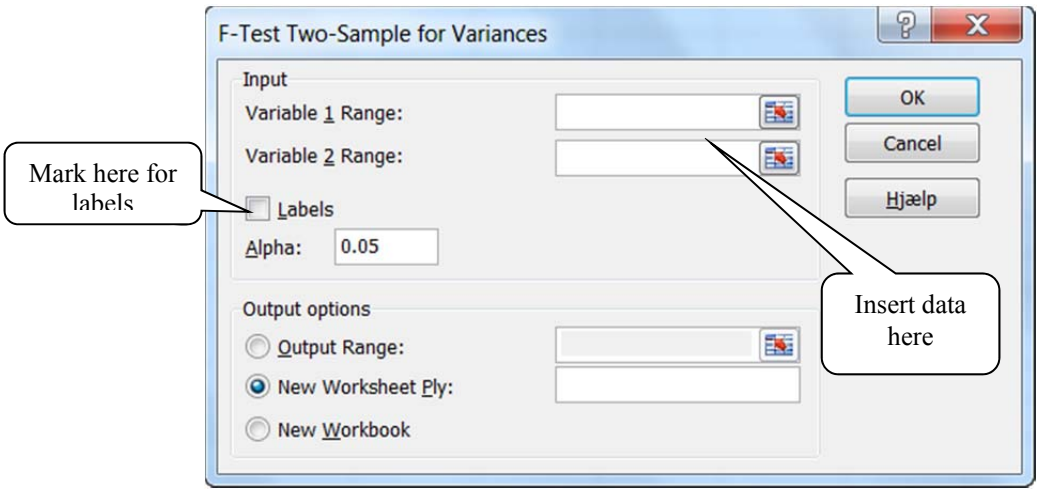
$$H_0: \frac{s_1^2}{s_2^2} = \frac{3.1}{1.7} = 1.82$$

As 1.82 < 3.06 the H_0 is accepted. The variances among the two dataset are similar. This is perhaps surprising. How can 3.1 be equal to 1.7? This is so because the datasets considered are very small and therefore rather uncertain. If data are drawn from a very large total population the variance may easily be very different.

When the number of observations increases the variances become more equal or homogenous. Observe from the illustration above that when $df_1 = df_2 = \infty$ then F = 1.00. The critical value can never be below 1. *So the largest variance should always be in the numerator of the tester.*

In **Megastat** mark for *test for equality of variances* as shown on the screenshot page 6, and the test will be performed as a part of the t-test.

In **Excel**, there is a function for the F-test for variance. Type **Data | Data analyses | F-test: Two samples for variances**, and obtain the menu on shown on the screenshot. Then mark for the two data sets, and click on "OK".



Finally, the F-test for equal variance can be calculated by use of the **pocket calculator.** Tast STAT → TESTS → "E: 2-SampFtest" → STAT → (NOTICE here we have to insert the *standard deviations*) so insert s_1, s_2, n_1, n_2 and set the test up. It should be $> \sigma_2$ if data has been set up correctly → CALCULATE, and the results will appear on the screen. The output will give the p-value. In the case it is equal to 0.14. This is consistent with the conclusion above that H_0 is accepted.

5. Worked Examples

1. The Effect of the Different type of Engine Oil for a Car[1]

Let us see the t-test and the F-test for variance in action. A popular producer of cars has sold many cars in Denmark with a particularly diesel engine. This engine has a service interval equal to 30,000 kilometers.

However, in the fall of 2007 the producer announced a change in the service interval from 30,000 kilometers to 20,000 kilometers. This was not good news for the consumers, because a shorter service interval means higher operating costs. Consequently, there have been claims from consumer groups over the issue.

A service center suggests a customer, that a different type of motor oil can be used. This oil is more efficient, and the motor is claimed to use less gasoline. This may compensate for the increase in service costs, because the car should be able to run longer per liter gasoline.

The customer has examined this issue, and collected statistics on the use of gasoline per liter before and after the different motor oil was filled on the motor. The table below gives the relevant statistics.

Before	16.6	17.5	16.8	17.2	15.1	16.1	15.8	16.3	16.1	15.8	16.3	17.2

After	17.8	17.5	16.9	18.1	17.8	16.6	17.2	16.9	17.5	19.2	17.2	18.1	18.8	16.3

A. Calculate mean and standard deviation for the two data set
B. Examine a hypothesis stating that the new oil has improved the number of driven kilometers per liter gasoline
C. Examine that the assumption of equal variance is fulfilled

The data for this example can also be found in Excel format on the file **BAINT 07 Two sample test case.xls.**

A)
Calculate mean and standard deviation for the two data sets

Before (B) $\bar{X}_B = \frac{\sum_{i=1}^{n_B} x_{B.i}}{n_B} = \frac{196.8}{12} = 16.40$ km per litre gasoline

[1] This exercise has been taken from the 2009 BA exam in statistics. Data are from the real life. The car producer is Peugeot and the engine is the 1.6 HDI diesel.

$$s_B = \sqrt{\frac{\sum_{i=1}^{n_B}(x_{B.i} - 16.40)^2}{n_B - 1}} = 0.6941$$

These results can easily be obtained by use of a calculator.

For the standard deviation use for example the formula:

$$s_B = \sqrt{\frac{1}{n_B - 1}\left[\sum_{i=1}^{n_B} x_{B.i}^2 - \frac{(\sum_{i=1}^{n_B} x_{B.i})^2}{n_B}\right]} = \sqrt{\frac{1}{12-1}\left[3{,}232.82 - \frac{(196.8)^2}{12}\right]} = 0.6941$$

After (A): $\overline{X}_A = \frac{\sum_{i=1}^{n_A} xA._i}{n_A} = \frac{245.9}{14} = 17.56$ km per litre gasoline

$$s_A = \sqrt{\frac{\sum_{i=1}^{n_A}(x_{A.i} - 17.56)^2}{n_A - 1}} = 0.8120$$

Alternatively:

$$s_A = \sqrt{\frac{1}{n_A - 1}\left[\sum_{i=1}^{n_A} x_{A.i}^2 - \frac{(\sum_{i=1}^{n_A} x_{A.i})^2}{n_A}\right]} = \sqrt{\frac{1}{14-1}\left[4{,}327.63 - \frac{(245.9)^2}{14}\right]} = 0.8120$$

B)
Examine a hypothesis stating that the new oil has improved the number of driven kilometers per liter gasoline

We can use the test for equal variance because we tested it for this above and accepted H_0.

The formulation of the problem is a little bit tricky, because the test should be set up as a one-sided test.

Hypotheses:	H_0: $\mu_A \leq \mu_B$	No effect of new oil
	H_1: $\mu_A > \mu_B$	New oil improves efficiency

I have set it up as a one-sided test. Alternatively, a two-sided test could have been applied. The ”pooled” variance can be calculated as:

$$S_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(14-1)(0.8120)^2 + (12-1)(0.6941)^2}{14 + 12 - 2} = \frac{8.5715 + 5.2995}{24} = 0.5779$$

Then we calculate the tester:

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{17.56 - 16.40}{\sqrt{0.5779\left(\frac{1}{14} + \frac{1}{12}\right)}} = \frac{1.16}{0.2990} = 3.89$$

Degrees of freedom equals: $Df = n_A + n_B - 2 = 14 + 12 - 2 = 24$

Two-sided test: $t_{0.025(24)} = 2.064 < 3.56$ *so* H_1 *is accepted*
One-sided test: $t_{0.050(24)} = 1.711 < 3.56$ *so* H_1 *is accepted*

So we find that the new oil improve the efficiency of the car (regardless of set-up).

This task can also be undertaken in **Excel**. Use: **Data | data analysis | T-test: Two-sample assuming equal variances**

t-Test: Two-Sample Assuming Equal Variances

	After	***Before***
Mean	17.56	16.40
Variance	0.66	0.48
Observations	14	12
Pooled Variance	0.58	
Hypothesized Mean Difference	0	
df	24	
t Stat	3.89	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.71	
P(T<=t) two-tail	0.00	
t Critical two-tail	2.06	

C)
Examine that the assumption of equal variance is fulfilled

We have two samples of equal side with $n_B = 12$ and $n_A = 14$.

We use the method outlined above, and set up the following:

Hypotheses: H_0: $\sigma_A^2 = \sigma_B^2$ (the two series has equal variance)
H_1: $\sigma_A^2 > \sigma_B^2$ (variance *after* is the largest)

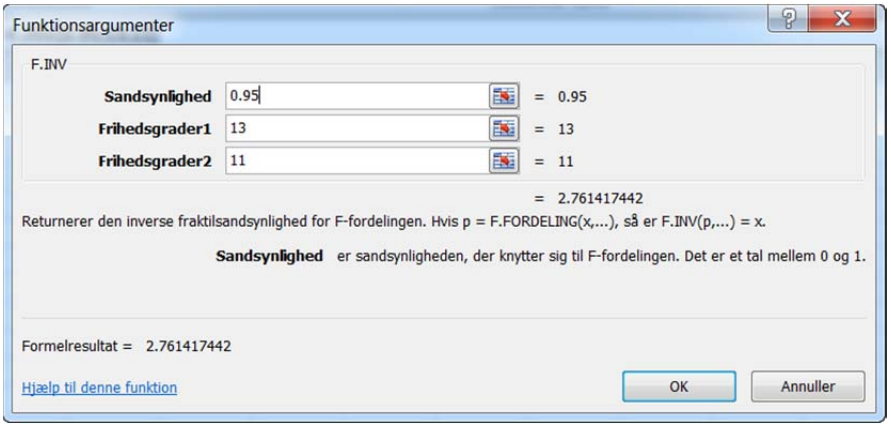
Test:

$$F_{\alpha(n_A-1)(n_B-1)} = \frac{s_A^2}{s_B^2} = \frac{(0.8120)^2}{(0.6941)^2} = \frac{0.6593}{0.4818} = 1.37$$

Normally the largest value is placed above!

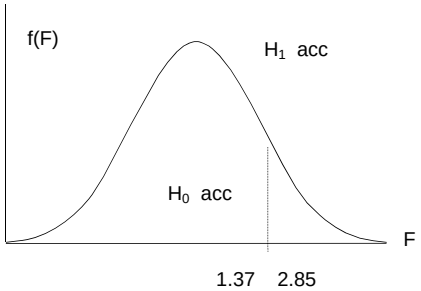
I have performed a one-sided test and assumed that α=0.05. In **Statistics Tables** the critical F-value can be found as $F_{0.05(13.11)} = 2.85$ (Here I have used $F_{0.05(15.10)}$ as the most close value in the table)

The precise F-value can be found by use of Excel. Use: **Formulas | Insert Function | FINV** and obtain:



As 1.37 < 2.85 H_0 is accepted and the two variances are identical. The assumption for the test performed under question B is then fulfilled.

Illustration:



The test can also be undertaken in **Excel.** Use: **Data | data analysis | F-test: Two samples for variance.**

Obtain:

F-Test Two-Sample for Variances

	After	*Before*
Mean	17.56	16.40
Variance	0.66	0.48
Observations	14	12
df	13	11
F	1.37	
P(F<=f) one-tail	0.30	
F Critical one-tail	2,76	

Finally the same results can be found by use of **Megastat.** Use the box displayed on page 6 in these notes. The following output will appear (including the t-test as well as the F-test for equal variances):

Hypothesis Test: Independent Groups (t-test, pooled variance)

After	Before	
17.564	16.400	mean
0.812	0.694	std. dev.
14	12	n

24	df
1.1643	difference (After - Before)
0.5780	pooled variance
0.7603	pooled std. dev.
0.2991	standard error of difference
0	hypothesized difference
3.89	t
.0003	p-value (one-tailed, upper)

F-test for equality of variance

0.659	variance: After
0.482	variance: Before
1.37	F
.6095	p-value

2. Test for the Efficiency of Entrepreneurship

In set 2 of the notes to Statistics I, we considered the case of the drive-in facility of the bakery's in Haderslev and Aabenraa. The queue at the drive-in was modelled by use of a Poisson process. In this case we continue working with this problem, and extend the analysis to consider data from the Global Entrepreneurship Monitor (GEM). The test for comparison of two population proportions is used in the present case.

In the example newspaper described the opening of the new bakery in Aabenraa, and praised the entrepreneurship talent not only of the baker, but also in general of the people in the region of Southern Denmark.

According to the Global Entrepreneurship Monitor (GEM) statistics there were 4,528 small entrepreneurship firms in the region out of a total equal to 40,428 small firms in the region. In the rest of Denmark, these numbers amounted to 15,559 small entrepreneurship firms out of a total equivalent to 123,484 small firms.

Problem statement

Examine by use of an appropriate test if the share of entrepreneurship firms in the region of Southern Denmark is larger than the share in the rest of Denmark

This problem is about testing the hypothesis that the two proportions are equal.

Let us initially state the hypothesis:

H_0: $\hat{p}_{SD} - \hat{p}_{DK} \geq 0$ (the share of entrepreneurship firms is similar or larger in the region of South Denmark)

H_1: $\hat{p}_{SD} - \hat{p}_{DK} < 0$ (the share of entrepreneurship firms is smaller in the region of South Denmark)

A one-sided test is considered.

With subscript SD = region of South Denmark and DK = Denmark total. As nothing is assumed with regard to the level of significance we use a 5 % test so α=0.05. We now conduct the test along the procedure described in Bowerman.

The two shares can be found as $\hat{p}_{SD} = x_{SD}/n_{SD} = 4,528/40,428 = 0.1120$ and $\hat{p}_{DK} = x_{DK}/n_{DK} = 15,559/123,848 = 0.1256$.

Now we calculated the "pooled share" (this has to be used in order to calculate the "pooled variance"):

$$\hat{p} = \frac{x_{SD} + x_{DK}}{n_{SD} + n_{DK}} = \frac{4,528 + 15,559}{40,428 + 123,848} = \frac{20,087}{164,276} = 0.1223$$

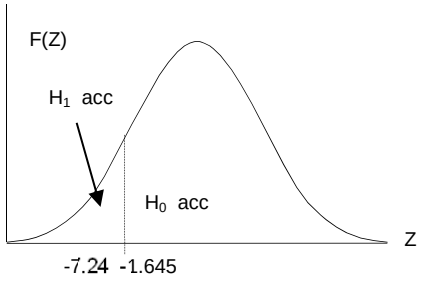
We have a very large data set, so data are normally distributed (Z-distribution). We then calculate the tester:

$$Z = \frac{\hat{p}_{SD} - \hat{p}_{DK}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{SD}} + \frac{1}{n_{DK}}\right)}} = \frac{0.1120 - 0.1256}{\sqrt{0.1223(1-0.1223)\left(\frac{1}{40{,}428} + \frac{1}{123{,}848}\right)}} = \frac{-0.0136}{0.001877} = -7.24$$

With $\alpha = 0.05$ we have the corresponding value of Z is 1.645.

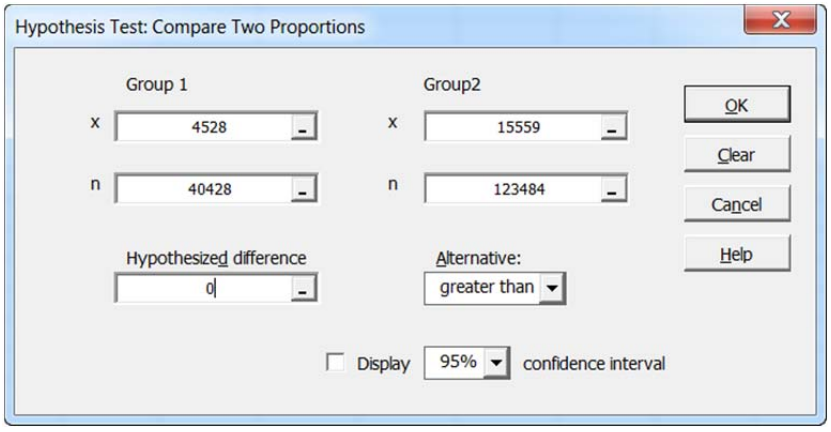
As $-1.645 > -7.24$ we *accept* H_1. So the share of innovative firm in the region of South Jutland is smaller than for the rest of Denmark.

Illustration:



We could alternatively have provided a two-sided test. In this case the critical value is equal to –1.96. We reach then the same conclusion.

On the computer open **Megastat** and select *Hypothesis tests* then select: *Compare two independent proportions*. The following dialog box will appear:



I have filled out with the data here. Click on OK and get the output:

Hypothesis test for two independent proportions

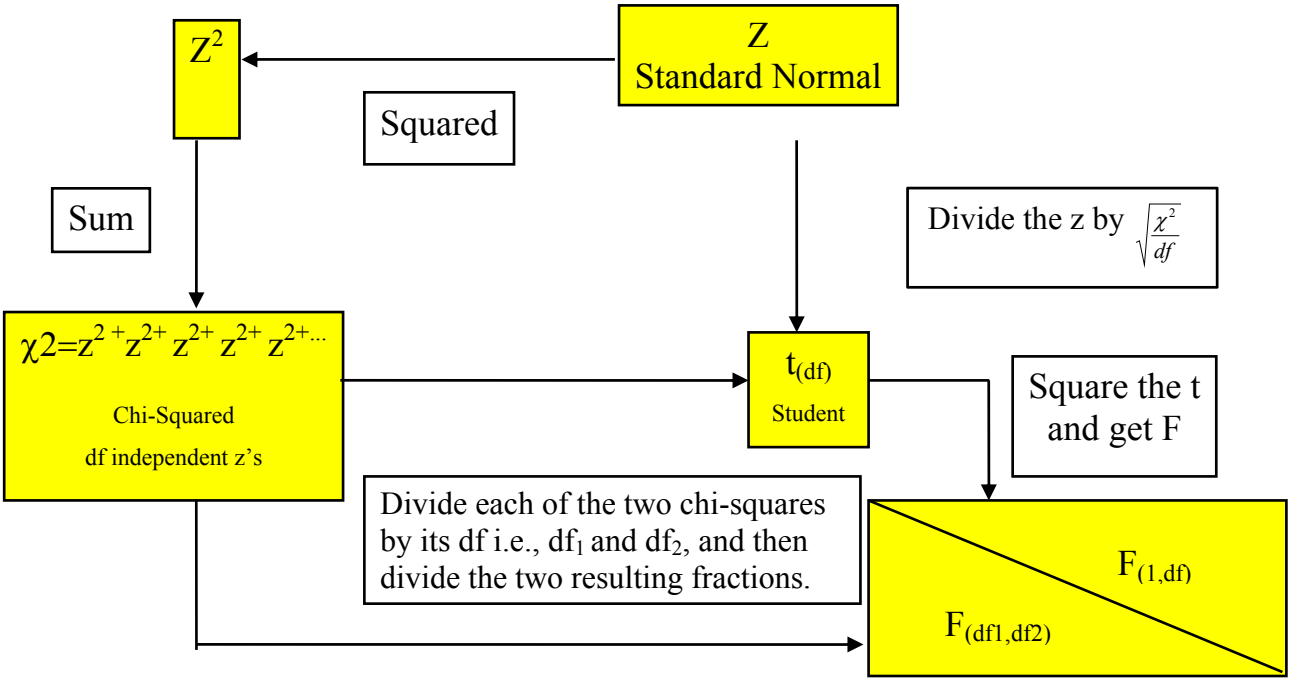
p1	*p2*	p_c	
0.112002	0.126	0.122547	p (as decimal)
0.112002	0.126	0.122547	p (as fraction)
4528	15559	20087	X
40428	123484	163912	n
		-0.014	difference
		0	hypothesized difference
		0.001879	std. error
		-7.45006	z

(The numbers are a little different for the z-value but the outcome is the same).

6. Relations among Probability Distributions Useful in Statistics

During note sets 6 and 7 several new distributions have emerged. Common is that they all are based on various manipulations of the Normal Distribution.

The illustration below provides an overview of the relations among them.



Nils Karl Sørensen ©
Statistics II, BA-INT

Set 8: One-Way Analysis of Variance (ANOVA)

by Nils Karl Sørensen

Outline

page

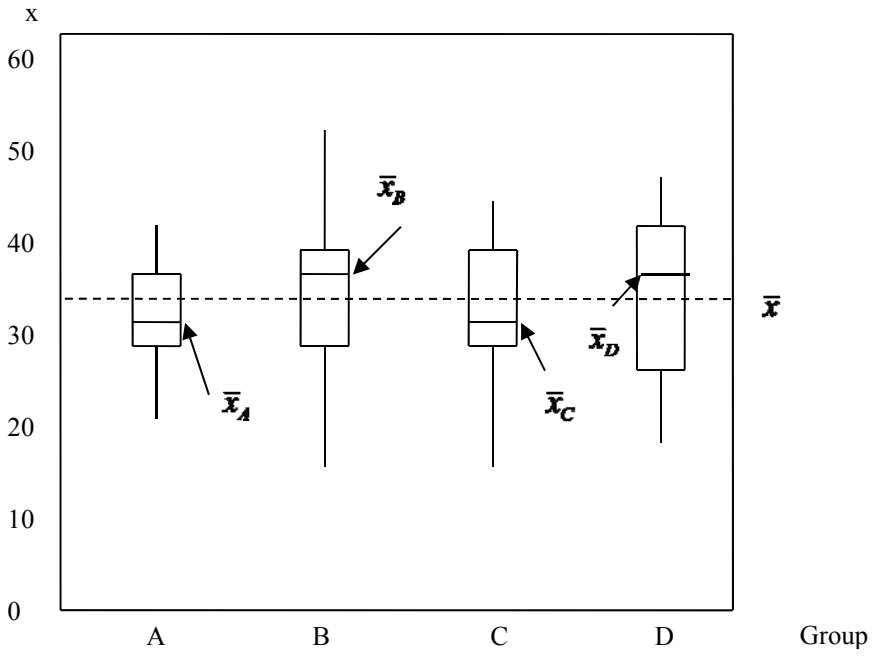
1. What is ANOVA? Setting up the Assumptions

In the two previous sets of notes, we have worked with hypothesis test with regard to a single or two data series. In this set of notes we consider a general situation where we want to compare p groups for equal mean. For example we want to compare sales performance in p different sales districts or comparing p years.

The method used to deal with this issue is called *ANOVA* or analysis of variance. The name may at first sound a little misleading because we are testing for equal mean. However, as it will be evident the variance component plays a very central role in this analysis.

Analysis of variance may be one-way or two-way in design. In this course we shall only deal with one-way analysis of variance. Two-way analysis or multivariate variance analysis is beyond the scope of this course, but is very frequently used in marketing research.

The diagram on the next page with 4 modified box-plots may help us in exploring the situation. For simplicity we consider a situation with 4 populations or groups labeled A, B, C and D, so $p = 4$. The samples do not have to be of equal size. The modification of the box-plot is that we have replaced the median in the diagram with the sample mean in order to explore the point of the analysis.



We want to test a hypothesis stating that the 4 means are equal against the alternative that at least one is different. The 4 samples may be turnovers in 4 different markets and we want to find the best market performance. That is the one with the highest mean. By inspection of the diagram we can observe that the samples not are equal with regard to range. Sample B and D clearly have a larger range. The means of these two groups are also higher than for A and C. However, is this difference significant? If we compute the overall mean $\bar{x}$ (shown by the dotted line in the diagram) then it will be somewhere in-between the 4 group means.

In order to proceed, we need some kind of a reference that relates the 4 groups to each other. This could be the overall mean. By use of the overall mean we can define two types of variation namely:

- Variation between groups: Here a given observation referees to the overall mean
- Variation within groups: Here a given observation referees to the group mean

The idea in ANOVA is to relate these two types of variation to each other and see if the mean difference if significantly different. In order to conduct the ANOVA we make the following assumptions.

Assumptions for ANOVA

1. Constant variance (homogeneity), i.e. $\sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$
2. Normality – all data are normally distributed
3. Independence

2. One-Way Analysis of Variance

For each of the p groups we can calculate the population (or sample) mean μ_i, $i = 1,2,...,p$. We may state the hypothesis for the ANOVA analysis as:

H_0: $\mu_1 = \mu_2 = ... = \mu_p$
H_1: Minimum one is different from the others

Each group has n_i observations. The groups may be of different size. The total number of observations is equal to $n = n_1+n_2+...+n_p$.

For a given observation j in group i called x_{ij} we can look at two types of variation namely the variation between groups called *SSTR* and the variation within groups called *SSE*. Finally, the total variation called *SST* can be computed as the sum of *SSTR* and *SSE*. Such variation is considered relative to the relevant mean. As our observation may be located above or below the mean the variation may be positive or negative. In order to avoid that positive and negative variation cancel out, we square all variation.

Let us summarize our findings on variation:

Total variation	= between groups	+ within groups
Sum Square Total	= Sum Square Treatment	+ Sum Square Error

Or:

$$SST = SSTR + SSE$$

On formula:

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Notice that the group mean $\bar{x}_i$ cancels out. The expression may seem complicated, but move back to illustration above and try to work it out. Then things should (hopefully) be clearer.

Finally what is "treatment" and "error"? We shall return to this issue in the note set 10 on regression. Very briefly *treatment* is the explanation of a given *response variable*. So this is how good the p variables "treat" or explains each other (between the groups). Then the remaining variation is within groups. So it does not contribute to explanation between groups. Therefore it has no impact on the total explanation, and it is as a result treated as an error.

The ANOVA-table and Conducting the Test

We are now in a position to conduct a one-way ANOVA analysis. We have decomposed our variation into variation among groups and the variation between groups. We started out with *p* classifications of variation, and this has been reduced to two types of variation in form of a squared sum.

When we worked with the test for variance above, we developed a tester based on the ratio among the two variances, and used the F-distribution as statistical reference. Here we proceed in a similar way. This is actually the reason why this method is named *analysis of variance.*

In testing for equal variance among two data sets we used the ratio of the variances, so we had normalized the tester by the number of observations in each sample minus one ($df = n_i - 1$. We divide each variance by $n_i - 1,\ i=1,2$).

Here we proceed in a similar way, but now we have *p* groups. We divide the sum of squares by the degrees of freedom, and then obtaining the *mean square treatment of regression (MSR)* and the *mean square error (MSE):*

Between groups: $MSR = \frac{SSTR}{p-1}$ (for the *p* groups minus the overall mean)

Within groups: $MSE = \frac{SSE}{n-p}$ (for the *n* minus the *p* groups)

Finally we can obtain a tester following the F-distribution by division of the two mean squares:

$$F = \frac{MSR}{MSE} \qquad \text{with } df=(p-1);(n-p)$$

All these calculations can by summarized into the *ANOVA-table:*

Variation	*Squared sum (SS)*	*Degrees of freedom (df)*	*Mean square (MS)*	*F-value*
Between groups	$SSTR = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2$	p − 1	MSR = SSTR/(p−1)	$F = \frac{MSR}{MSE}$
Within groups	$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	n − p	MSE = SSE/ (n−p)	
Total	$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$	n − 1		

The "F-value" is then F-distributed with degrees of freedom equal to (p−1);(n−p).

The total number of observations is equal to $n = n_1+n_2+\ldots+n_p$.

Observe how nice things turn out. Everything is related to each other in the ANOVA-table. For the total the number of degrees of freedom is equal to $n-1$, because p cancels out.

Supplemental or Post-Hoc Analysis

So now we are happy? Well, if we have accepted the alternative hypothesis then things may not be so easy. We have to find the group (or groups) that have a significantly different mean than the others. We can setup *pair wise confidence intervals by groups* in order examine this issue.

- For difference among two groups i and h:

$$\left[(\bar{x}_i - \bar{x}_h) \pm t_{\alpha/2} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_h}\right)} \right] \qquad \text{where } df = n-p$$

- Point estimate of group i mean

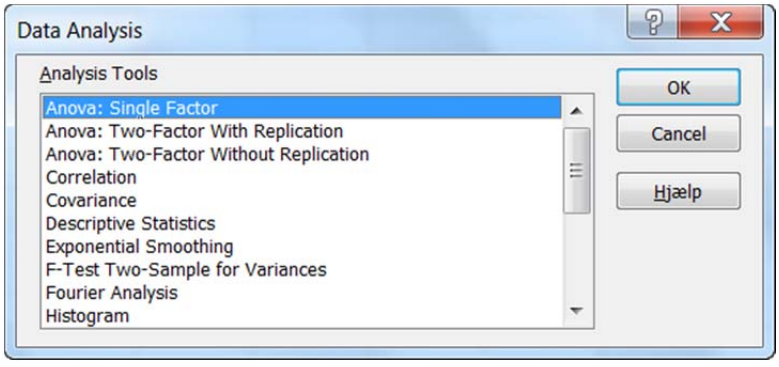
$$\left[\bar{x}_i \pm t_{\alpha/2} \sqrt{\frac{MSE}{n_i}} \right] \qquad \text{where } df = n-p$$

We compare the par wise or point confidence intervals, and examine if some of the intervals are located inside or outside the interval range of the other groups. In case the range for a specific variable is outside the range of the confidence interval for the other variables then H_1 is accepted. This mean of this group is then significantly different from the means of the other groups.

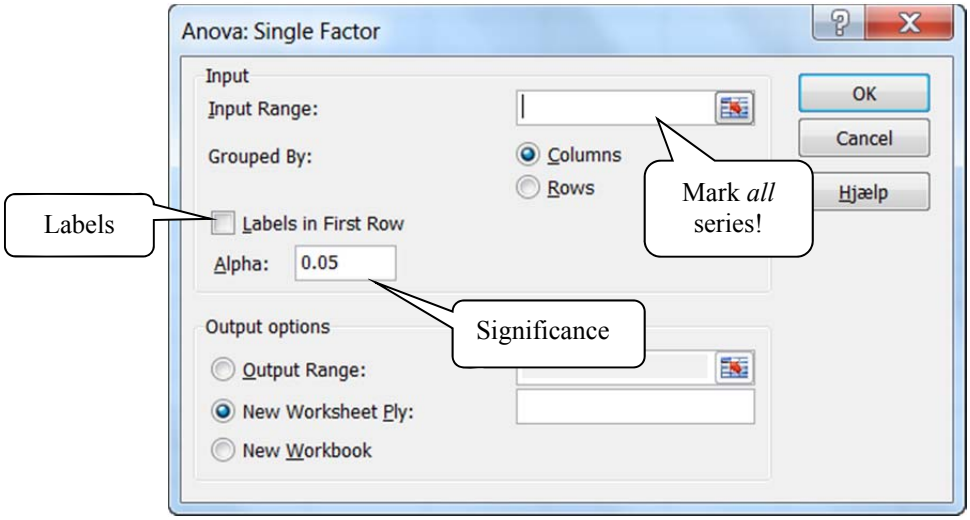
ANOVA by Excel

We can perform the ANOVA analysis by use of Excel. We have to perform the supplemental analysis manually. In Excel choose:

1. **Data | Data Analysis | Anova: Single Factor**



2. We obtain the following dialog box



3. Mark also ”labels” (DK: ”etiketter”) and select the critical value ”alpha” (normally 0.05. That is a 95% level of significance with a one-sided test).

ANOVA on a Pocket Calculator

On the **TI-84 pocket calculator** the first task is to plot in the *p* dataset in the data register. If we fore example have 3 data series then use *L1, L2* and *L3*. Then select: STAT → TESTS → H: ANOVA(→ ENTER.

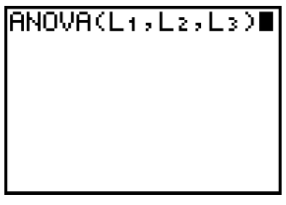
The format is ANOVA(*L1,L2,L3*) → ENTER. See the example taken from the TI-84 manual:

ANOVA(*list1*,*list2*[,...,*list20*]**)**

In the example:

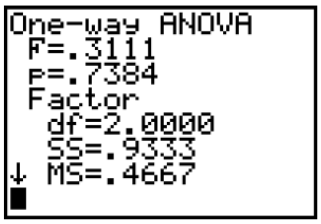
L1={7 4 6 6 5}
L2={6 5 5 8 7}
L3={4 7 6 7 6}

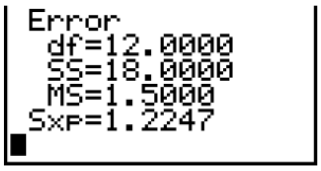
Input:



The following will appear on the screen of the calculator:

Calculated results:

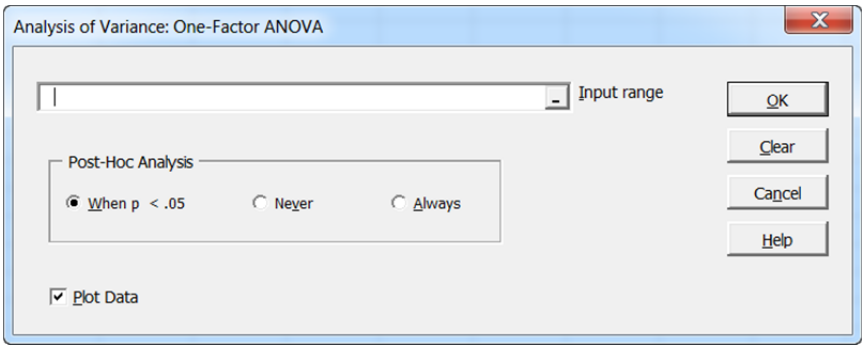




Notice that **SS** is the sum of square and **MS** is mean square. Observe that the output shown on the screen of the calculator constitutes the elements of the ANOVA table.

ANOVA by Megastat

It is also very easy to conduct the ANOVA-analysis by use of Megastat. Open Megastat and select Analysis of Variance and then One Factor ANOVA. Get:



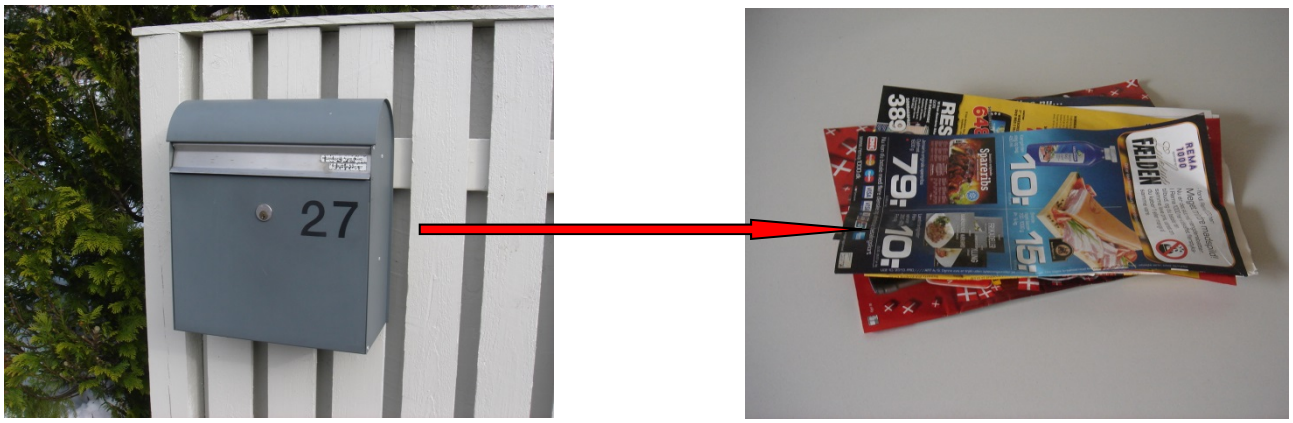
Load all data into the input range and click OK. Observe two points. First, Megastat performs a very nice plot of the data. See next Section. Second, Megastat also performs a post-hoc analysis and calculates the so-called Tukey analysis. This topic is not a part of the course, so use the approach with the confidence intervals shown above.

4. Example of ANOVA: The Case of Seasonality

Let us illustrate the ANOVA analysis by use of an example putting the focus on seasonality.

Studies of seasonal patterns in quarterly data for many national economic variables in a broad section of OECD countries show a strikingly identical pattern[1]. The first and third quarters have low economic activity, whereas the second (before the holidays) and fourth quarters (before Christmas) are characterised by high demand and high economic activity. These circumstances may well influence the amount of advertising that goes out to households in the form of distributed printed advertisements on weekends.

How can we use ANOVA to investigate for seasonality in the form as claimed? Each week in 2004 the amount of printed advertisements received in my own mailbox in Haderslev was weighted on a scale. If the economic activity is high then many advertises should be received and vice versa. The weight of advertises per week should then over a year reflect the seasonal pattern of economic activity claimed. The scene of the crime and advertise sample is shown below



The 52 weeks forming a year can now be divided into 4 quarters of 13 weeks each. We can use this material with 4 groups, and examine the seasonal behaviour relative to the pattern postulated above.

Let us label a quarter by q. For each quarter we can calculate the mean weekly weight of advertises and compare for mean. If we do not have seasonal variation then the means should be equal

[1] See Svend Hylleberg, Nils Karl Sørensen and Clara Jørgensen "Seasonality in Macroeconomic Time Series". *Studies of Empirical Economics 18,* pages 321−335, 1993.

We can set up the following hypothesis:

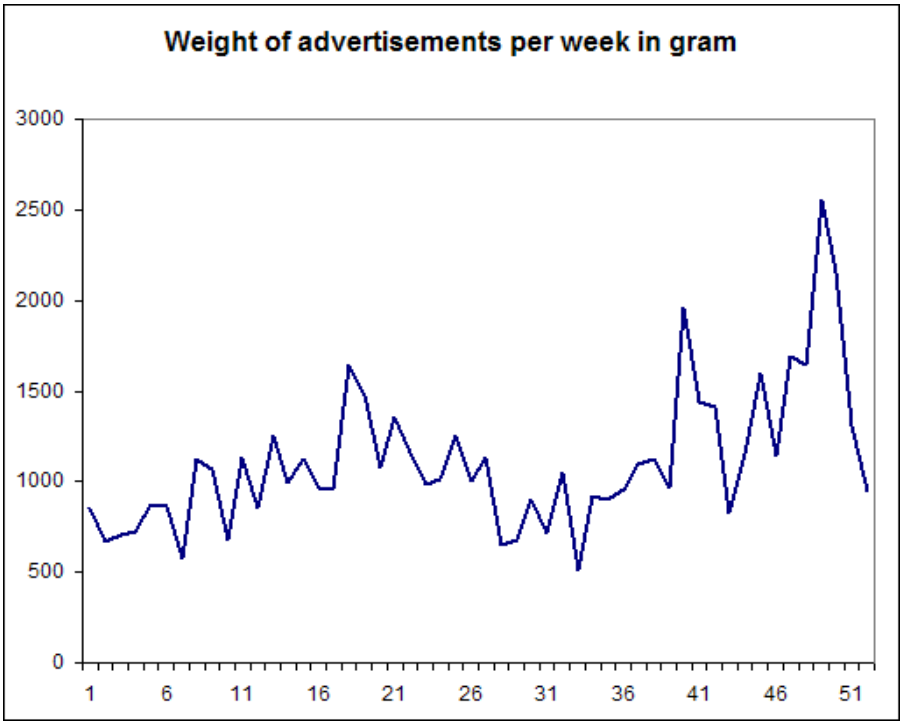
H_0: $\mu_{q.1} = \mu_{q.2} = \mu_{q.3} = \mu_{q.4}$ (no seasonality)
H_1: Minimum one mean is different (seasonal pattern)

Data for this analysis can be found in the file **BAINT 08 Example ANOVA on Seasonality.xls.** Let us first look at the data and a graph:

Weight in gram per week

Week no.	1 q	2 q	3 q	4 q
1	854	996	1,135	1,960
2	670	1,124	647	1,436
3	709	965	679	1,410
4	723	960	900	824
5	874	1,640	718	1,151
6	863	1,480	1,048	1,599
7	577	1,075	510	1,139
8	1,126	1,352	923	1,691
9	1,070	1,162	903	1,640
10	675	983	956	2,557
11	1,130	1,009	1,098	2,158
12	852	1,258	1,121	1,305
13	1,256	1,007	962	944

Diagram in time



Inspection of the graph confirms some of our claims. For example the amount of advertises increases at the end of the year ultimo November and primo December (Christmas effect). It also seems evident that the amount of advertises is low in weeks 26-31 i.e. the holidays season.

We can use **Excel**. We apply the procedure outlined above and obtain the output:

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
1 quarter	13	11,379	875.31	44,484.56
2 quarter	13	15,011	1,154.69	47,490.56
3 quarter	13	11,600	892.31	39,166.40
4 quarter	13	19,814	1,524.15	240,713.81

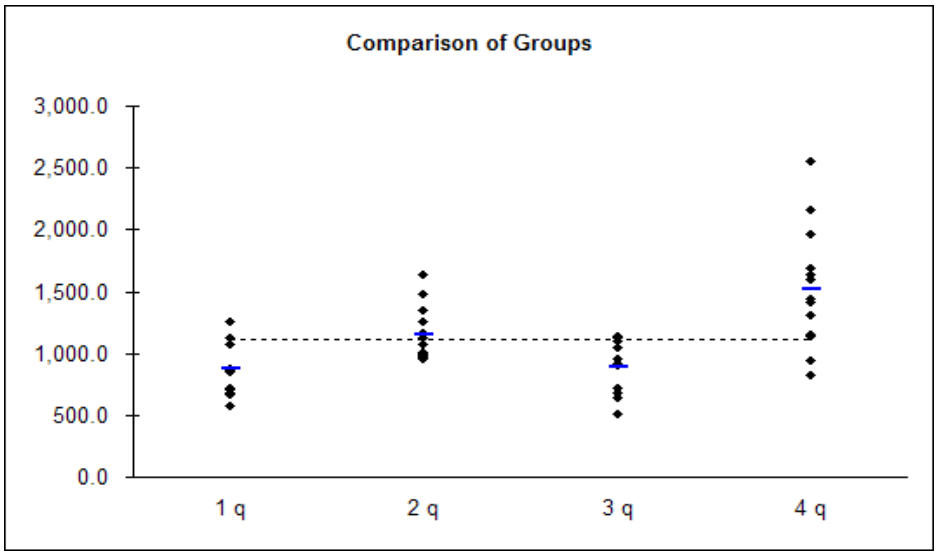
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3,587,750	3	1,195,916.77	12.86	0.00	2.80
Within Groups	4,462,264	48	92,963.83			
Total	8,050,014	51				

We observe the F-value of 12.86. The critical value at $\alpha = 0.05$ is $F_{0.05}(3,48) = 2.81$[2]. This is obtained from **Statistics Tables.**

On this background we reject H_0 and *accept* H_1. Consequently a seasonal pattern is observed.

We can obtain further insight into the seasonal pattern from the graph below



[2] I use for (3,45) due to lack of degrees of freedom in the table in **Statistics Tables**.

The graph shows observations and means by quarter. It is evident that:

- The lowest amount of advertises is in the first and third quarter
- The highest amount is in the fourth quarter

Supplementary or Post-hoc Analysis
In order to verify the seasonal pattern we set up 95 % confidence intervals for each quarter and compare. We use the *MSE* from the ANOVA table, whereas the averages can be found in summary statistics. We obtain:

$$\overline{X}_{q.1} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_1}} \Rightarrow 875.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 875.31 \pm 2.01(84.56) \Rightarrow [705.34\,;1045.28]$$

$$\overline{X}_{q.2} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_2}} \Rightarrow 1154.69 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 1154.69 \pm 2.01(84.56) \Rightarrow [984.39\,;1324.66]$$

$$\overline{X}_{q.3} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_3}} \Rightarrow 892.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 892.31 \pm 2.01(84.56) \Rightarrow [722.34\,;1062.28]$$

$$\overline{X}_{q.4} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_4}} \Rightarrow 875.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 1524.15 \pm 2.01(84.56) \Rightarrow [1354.18\,;1694.12]$$

There is no overlap between the fourth quarter and the others. So it is significantly different at the 95 % level. We cannot identify differences among the other quarters.

Manual Calculation of the F-tester in the ANOVA-table
If you not are lucky and have Excel installed, Megastat installed or a pocket calculator then a manual calculation of the tester has to be performed.☹

In order to undertake this calculation we need information about:

- The mean by *p* groups $\bar{x}_i$
- The total number of observations by group n_i
- The sample variance by group s_i^2

Assume that this information is present, for example in the form given in the summary output provided in the Excel output. No more information is available, and the ANOVA-table has to be constructed manually.

The following information is available:

Groups	*Count*	*Sum*	*Average*	*Variance*
1 Quarter	13	11379	875.31	44484.56
2 Quarter	13	15011	1154.69	47490.56
3 Quarter	13	11600	892.31	39166.40
4 Quarter	13	19814	1524.15	240713.81

To calculate the F-tester, the components of the ANOVA-table have to be calculated. Initially, the grand mean $\bar{x}$ has to be found. It is defined as a the weighted mean of the mean by groups:

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i \bar{x}_i}{n} = \frac{13 \times 875.31 + 13 \times 1154.69 + 13 \times 892.31 + 13 \times 1524.15}{52} = \frac{57803.98}{52} = 1111.62$$

Next, the variation between the groups has to be calculated given as $SSTR = \sum_{i=1}^{p} n_i (\bar{x}_i - \bar{x})^2$

$$\begin{aligned} SSTR = \sum_{i=1}^{p} n_i (\bar{x}_i - \bar{x})^2 &= 13 \times (875.31 - 1111.62)^2 + 13 \times (1154.69 - 1111.62)^2 \\ &+ 13 \times (892.31 - 1111.62)^2 + 13 \times (1524.15 - 1111.62)^2 \\ &= 725951.41 + 24115.32 + 625259.39 + 2212353.01 \\ &= 3587679.13 \end{aligned}$$

(13 can be set outside the parenthesis). The value found is very close to the one in the ANOVA-.table calculated by use of Excel.

The variation within the groups is defined as $SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$. Look carefully at this expression. For each group the square root of the distance between each observation and the mean by group has to be calculated. This is similar to the way that the variance is calculated. When computing the variance the sum is divided by ($n_i - 1$). So for each group it must be valid that $SSE_i = \sum_{i=1}^{n_i} (x_i = \bar{x}_i)^2 = (n_i - 1) \times s_i^2$. Both the observations and the variance are known so it is possible to calculate *SSE* as:

$$\begin{aligned} SSE = \sum_{i=1}^{p} (n_i - 1) s_i^2 &= (13 - 1) \times 44484.56 + (13 - 1) \times 47490.56 + (13 - 1) \times 39166.40 \\ &+ (13 - 1) \times 240713.81 \\ &= 533814.72 + 569886.72 + 469996.80 + 2888565.72 \\ &= 4462263.96 \end{aligned}$$

(It is more easy to put (13−1) outside the parenthesis). The value found is identical to the one found by use of Excel in the ANOVA-table.

Finally, the F-value has to be found. By use of the methods outlined above it is obtained that:

$$F = \frac{MSR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{3587679.13/(4-1)}{4462263.96/(52-4)} = \frac{1195893.04}{92963.83} = 12.86$$

This is similar to the value obtained by use of Excel.

Nils Karl Sørensen ©
Statistics II, BA-INT

Set 9: χ^2-test for Independence and Contingency Tables

by Nils Karl Sørensen

Outline

page

1. Introduction

Until this point, we have always assumed that we know the underlying sample distribution of the data set that we are dealing with. The sample distribution may follow for example a Normal distribution or a student t distribution if the sample is small.

However, this is not always the case. In such a situation, we perform a chi-square test. This is frequently called a *distribution-free test.* The test is very easy to perform, and it is widely used especially within marketing and medicine. There are two types of chi-tests, namely the *Goodness-of-fit Test* and the test for *independence.* We shall consider them by turn.

2. Chi-Square Goodness of Fit Tests

The data used in the Chi-Square tests are *enumerative.* The data are counts or frequencies. For example, n observations may be divided into C groups or categories by for example gender, age, education, preferences etc. This design is a perfect match to for examples questionnaires.

A *goodness of fit test* is a statistical test of how well our data support an assumption about the distribution of a population or random variable of interest. The test determines how well an assumed distribution fits the data.

The point is departure is a situation with k outcomes or classes. Each outcome has a probability equal to $p_1, p_2, \ldots, p_k$ and a sample of n observations each divided into the k outcomes or classes. We assume that the k probabilities are fixed falling into $i = 1,2, \ldots, k$ cells.

In order to set up chi-square analyses consider the following steps (very close to the procedure outlined in note set 6):

1. We hypothesize about a population by stating the null and alternative hypotheses.
2. We compute frequencies of occurrence of certain events that we expect under the null hypothesis. These give us the *expected* counts of data points in different cells.
3. We note the *observed* counts of data points falling in the different cells.
4. We consider the difference between the observed (O) and the expected (E). This difference leads us to a computed value of the chi-square statistic.
5. We compare the value of the statistic with critical points of the chi-square distribution and accept or reject the null.

Let us now state the hypothesis:

H_0: The data set can be approximated by the given reference
H_1: The data set can *not* be approximated by the given reference

If we are specific and relate to the k outcomes the hypotheses can be written as:

H_0: The probabilities of occurrence of events $E_1, E_2, \ldots, E_k$ are given by the specified probabilities $p_1, p_2, \ldots, p_k$.
H_1: The probabilities of the k events are *not* the p_i stated in the null hypothesis

The next step is to compute the expected values. This is undertaken by multiplying the probability of the event by the total number of observations. The expected count in cell i is then:

$$E_i = np_i$$

As the probability of an event always ranges between zero and one, the expected value is always a positive value. Having the observed as well as the expected values, we now need the tester. When the null is true, then our data fits the proposed distribution. This implies that the difference between the observed and the expected value is so small that the difference is only due to stochastic variation. As the difference between the observed value and the expected value can be positive as well as negative we square this difference. Further we divide by the expected value in order to normalize our tester. We use the expected value because it is calculated by the distribution stated under the null. Summing over all C cells or categories we obtain the tester as:

$$\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

With degrees of freedom equal to $df = C-1$.

Here O_i is the observed value for observation i, and E_i is the expected value. C is the number of cells (or categories) in the data set. The expected value is given as showed above by $E_i = np_i$ and $E_i \geq 5$. Finally, n is the number of observations in the data set considered, and p_i is the probability function for the given reference. This could be:

- Any data set we use
- A probability function imposed by a distribution for example a normal distribution, a Poisson distribution, a Binominal distribution, a uniform distribution etc.

Two points needs to be elaborated a little further. First, if the size of the expected value is smaller than five then the distance between the observed value and the expected value may be large. This implies that the chi-square statistic may be very large, and as a result we may have a bias towards accepting the alternative hypothesis. In order to avoid this situation cells with fewer expected observations should be grouped in such a manner that the expected value always exceeds five. As it will be evident from the exercises this may be a complicated puzzle because we also have to consider the object of our analysis.

Second, there is the issue of degrees of freedom. We lose one degree of freedom, because the observed values are taken from a sample with n observations. Then further restrictions may appear. For example if we have to calculate the sample mean $\overline{X}$ and the standard deviation s for the expected distribution then two more restrictions are imposed and $df = C-1-2 = C-3$.

Example: The Danish Grade Scale[1]

The Danish grade scale is very different from the German scale. The German scale has 10 different grades for the passed results, whereas the Danish scale only has 5 different grades.

The Danish grade scale has the grades −3, 00, 02, 4, 7, 10 and 12. From 02 and higher, the exam is passed. Observe the symmetry for the passed grades. From 02 to 4 there are two steps. From 4 to 7 there are 3 steps. From 7 to 10 there are 3 steps. Finally, from 10 to 12 there are 2 steps. The Danish grade scale can be seen from the table on the next page.

[1] The Danish 7-point grade scale was implemented in 2005. It has been developed in: Ministry of Education (2004): **Betænkning om indførelse af en ny karakterskala til erstatning af 13-skalaen afgivet af karakterkommissionen, november 2004.** Betænkning nr. 1453 (In Danish). Especially relevant is Chapter 8. I have been in dialog with the chairman of the commission who has confirmed that the interpretation of the scale as approximated by a Binominal distribution is correct.

If the Danish grade scale is centered around the grade 7 the following question can be raised: How large a share of the students that passes the exam should for example have 7? How about the distribution of the other grades?

At the exam it is normally required that if 50 percent of the answers are correct then the student pass the exam. At the exam there are two outcomes: passed or not passed. These two outcomes are equally likely. The statistical distribution underlying the Danish grade scale must then be a Binominal distribution with probability *p=0.5* and 5 outcomes for the passed grades.

A stochastic variables with *X = 0, 1, 2, 3* and *4* at fulfill this requirement. So it is assumed that *n=4*. Such a probability distribution can be found in **Statistics Tables** page 2. The distribution is evident from the third line in the table in the example below.

In order to illustrate consider an example of the distribution of the 92 grades for an exam in the course International Economics (VWL-III) that was held in February 2011 at the BA-INT study in Flensburg/Sønderborg. The distribution of grades on the Danish 7-point scale is given as:

Grades of passed (7-point DK scale)	2	4	7	10	12	Total
Frequency	10	26	33	19	4	92

This is the statistical material that also served as an example in the note set 1 on *Descriptive Statistics.*

The hypothesis for the test can be stated as:

H_0: The distribution of the grades at the VWL-III exam follows a Binominal distribution
H_1: The distribution of the grades at the VWL-III exam follows *not* a Binominal distribution

By point of departure in the frequency distribution above and the theoretical probabilities for the outcomes found in **Statistics Tables** page 2 the expected outcomes can be calculated. The expected values can be found in the fourth line of the table below. The value of the tester is calculated in the fifth line. The yellow shaded value is the chi-value found as the horizontal sum of the chi-values

Grades of passed (7-point DK scale)	2	4	7	10	12	Total
Observed (O_i)	10	26	33	19	4	92
Probability (p_i)	0.0625	0.250	0.375	0.250	0.0625	1.000
Expected ($E_i = np_i$)	5.75	23	34.5	23	5.75	92
$(O_i-E_i)^2/E_i$	3.141	0.391	0.065	0.696	0.533	4.826

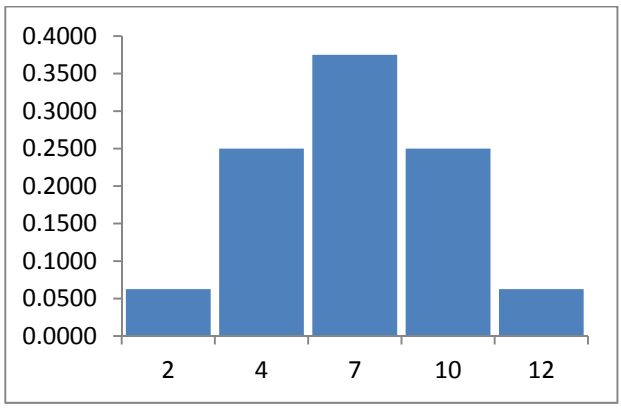
The tester can be found to be equal to $\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} = 4.826 \approx 4.83$

The number of degrees of freedom is $df = k-1 = 5-1 = 4$.

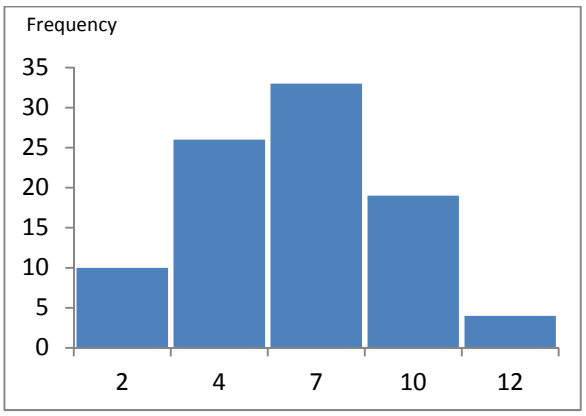
The critical value cab be found by use of **Statistics Tables** page 11 to equal $\chi^2_{(4)0.05} = 9.487$ assuming $\alpha = 0.05$. As $4.826 < 9.487$ H_0 is accepted. So the distribution of the grades can be said to be Binominal distributed with given probabilities.

The Goodness of fit test can also be computed by use of the pocket calculator or by use of Megastat. We shall consider this issue in Section 4 of these notes.

Finally, consider the illustration of the Binominal distribution to the left and the distribution of the grades to the right.



Binominal Distribution p=0.5 and n=4



Distribution of 92 grades in VWL-III

3. A Chi-Square Test for Independence – Contingency Tables

Recall the concept of *independence* that we discussed in the note set 5 to Statistics I. Two events are independent if the probability of their joint occurrence is equal to the product of their marginal (i.e. separate) probabilities. This was given as:

A and B are independent if $P(A \cap B) = P(A)P(B)$

We can tabulate two events by use of a *contingency table* – a table with cells corresponding to cross-classifications of attributes or events. For example we can tabulate gender and job performance or gender and preferences for different kind of soft drinks (diet or not diet). In market research studies such tables are referred to as *cross-tabs*. Normally a market questionnaire is designed in such a way that it is easy to set up such tables.

The contingency tables may have several rows R_i, $i=1,2,...,r$ and columns C_j, $j=1,2,...,c$ each responding to a classification category. A schematic example with $r = 4$ and $c = 5$ is shown below.

Layout of a Contingency Table

	First Classification Category					
Second Classification Category	**1**	**2**	**3**	**4**	**5**	**Total**
1	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	R_1
2	O_{21}	O_{22}	O_{23}	O_{24}	O_{25}	R_2
3	O_{31}	O_{32}	O_{33}	O_{34}	O_{35}	R_3
4	O_{41}	O_{42}	O_{43}	O_{44}	O_{45}	R_4
Total	C_1	C_2	C_3	C_4	C_5	n

Notice that n is the total number of observations.

We now want to examine for independence among the two variables. We proceed in a way similar to in the previous section. Let us first state the null and the alternative hypotheses.

H_0: The two classification variables are independent: $P(A \cap B) = P(A)P(B)$
H_1: The two classification variables are *not* independent: $P(A \cap B) \neq P(A)P(B)$

The definition of the tester is straightforward. The only difference is that the summation extends over all cells in the table. The chi-square test statistic for independence is so

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The degrees of freedom of the chi-square statistic are $df = (r-1)(c-1)$

All we need to do now is to find the expected cell counts E_{ij}. Here is where we use the assumption, that the two classification variables are independent. Look at a particular cell in row i and column j. In the previous Section the expected value was calculated as the sample size times the probability of the occurrence of the event signified by the particular cell. In the context of an $r \times c$ contingency table, the probability associated with cell (i,j) is $E_{ij} = nP(i \cap j)$. If we assume independence of the two classification variables, then event i and j are independent event, and by the law of independence among events, $P(i \cap j) = P(i)P(j)$.

From the row total, we can estimate the probability of event i as R_i/n. Similarly, we estimate the probability of event j by C_j/n. By substitution of the estimates of the marginal probabilities we get the following expression for the expected count in cell $(i, j) : E_{ij} = n(R_i / n)(C_j / n) = R_i C_j / n$.

So the expected count in cell *(i,j)* is $E_{ij} = \frac{R_i C_j}{n}$

In the analysis of 2×2 contingency tables, our chi-square statistic has *1 degree of freedom*. In such cases, the value of the tester is frequently "corrected" so that its discrete distribution will be better approximated by the *continuous* chi-squared distribution. This correction is called the *Yates correction*, and entails subtracting the number ½ from the absolute value of the difference between the observed and the expected counts before squaring them as required in the tester. The Yates corrected form for the tester is then

$$\text{Yates corrected } \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

Example of Test for Independence

Consider the contingency table below giving in the left panel the preferences by gender on two types of coca cola for 77 persons. We also considered this data set in the note set 5 on cross tabulation from the course in *Statistics I*. Let us investigate this relation further and provide a more precise answer than we were able to do in the *Statistics I* course. The hypothesis can be stated as:

H_0: The two variables are independent (no relation)
H_1: The two variables are not independent (relation)

What is the interpretation? If H_0 *not* can be accepted then there is a relation saying that for example females have a special preference of for example diet coke in the sample? This has a marketing implication!

The right panel shows the expected values calculated as for example $E_{11} = \frac{36 \times 33}{77} = 15.43$.

Sample of 77 persons Preferences for Coca-Cola (Coke) by Gender

Observed			
	Coke	Diet-coke	Total
Female	3	33	36
Male	30	11	41
Total	33	44	77

Expected			
	Coke	Diet-coke	Total
Female	15.43	20.57	36
Male	17.57	23.43	41
Total	33	44	77

Finally, we calculate the tester as:

$$\chi^2 = \frac{(3-15.43)^2}{15.43} + \frac{(33-20.57)^2}{20.57} + \frac{(30-17.57)^2}{17.57} + \frac{(11-23.43)^2}{23.43} = 10.01 + 7.51 + 8.79 + 6.59 = 32.90$$

Degrees of freedom are *df = (c−1)(r−1) = (2−1)(2−1) = 1*

Assuming α = 0.05 then $\chi^2_{(1)} = 3.84$. As 32.90 > 3.84 we cannot accept H_0. So a relation between gender and preferences for coke is observed. Females prefer diet-coke.

However, we are in the special case with *df = 1,* so let us finally calculate the Yates correction as:

$$\chi^2 = \frac{(|3-15.43|-0.5)^2}{15.43} + \frac{(|33-20.57|-0.5)^2}{20.57} + \frac{(|30-17.57|-0.5)^2}{17.57} + \frac{(|11-23.43|-0.5)^2}{23.43}$$
$$= 9.22 + 6.92 + 8.10 + 6.07 = 30.31$$

As evident the value of the tester decreases correcting the bias towards accepting the alternative. However, as 30.31 > 3.81 the conclusion obtained above is confirmed.

Cramer's V

The power of the test for independence can be expressed by *Cramer's V*. This is a measure very similar to the covariance or correlation. *Cramer's V* ranges between 0 and 1. A high value of *Cramer's V* indicates a high degree of dependence among the two variables being examined. *Cramer's V* is given as:

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1; c-1)}}$$

Here min(r−1;c−1) indicates that the minimum value of the row or he column minus one should be used

Example

In the example above *Cramer's V* can be calculated. In that case it is observed that *r = c = 2.*

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1; c-1)}} = \sqrt{\frac{32.90}{77 \times (2-1)}} = 0.65$$

This is a rather strong relation.

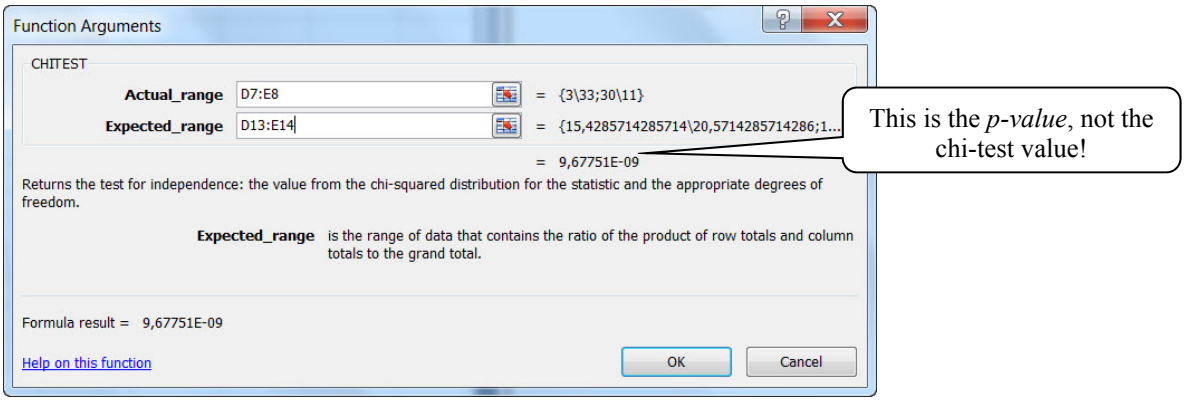
4. Chi-Square Tests in Excel, Megastat and on the Pocket Calculator

The goodness-of-fit test as well as the test for independence chi-squared tests can easily be performed by either Excel or Megastat. I advocate the use of Megastat. In Excel you have to undertake several of the calculations by yourself. In Megastat you obtain all calculations by marking in a dialog box, and the output is very easy to interpret.

Goodness of fit test and chi-Square Test in Excel

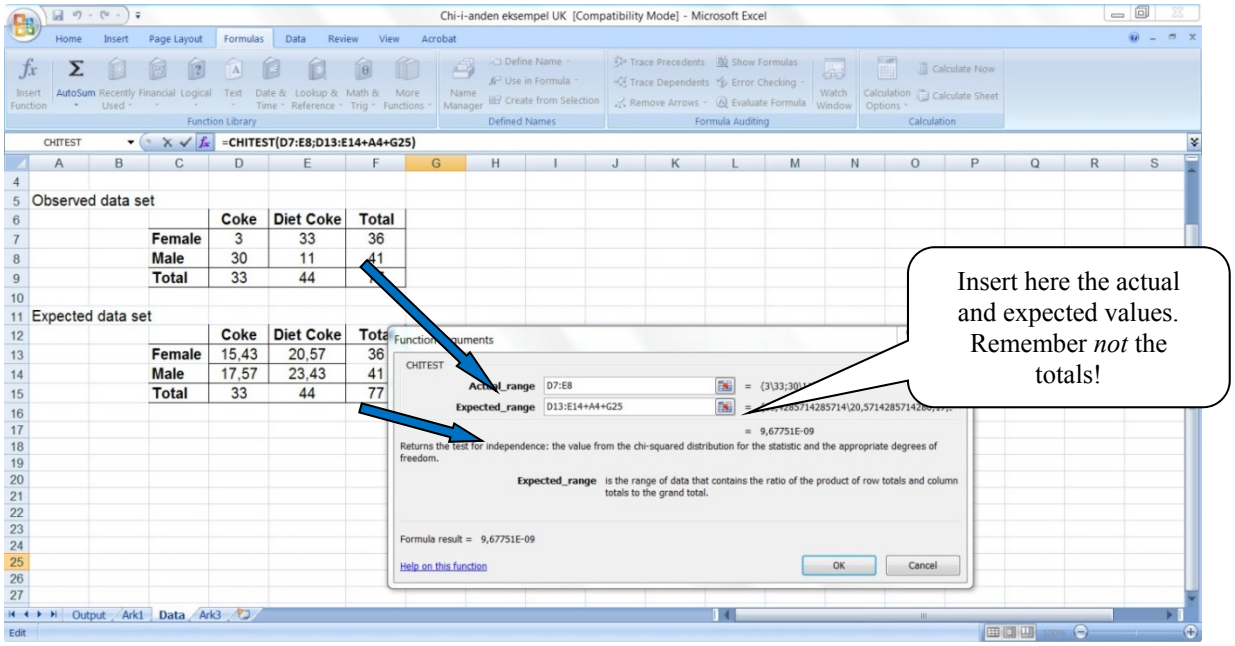
1. Select **Formlas | insert functions | statistical | chitest**.

Then we obtain the screenshot shown next page. Notice that we have to calculate the expected values manually by Excel, so we can mark the array. Further, we only obtain the critical value, not the value of the tester. The procedure is similar with regard to the goodness of fit test.



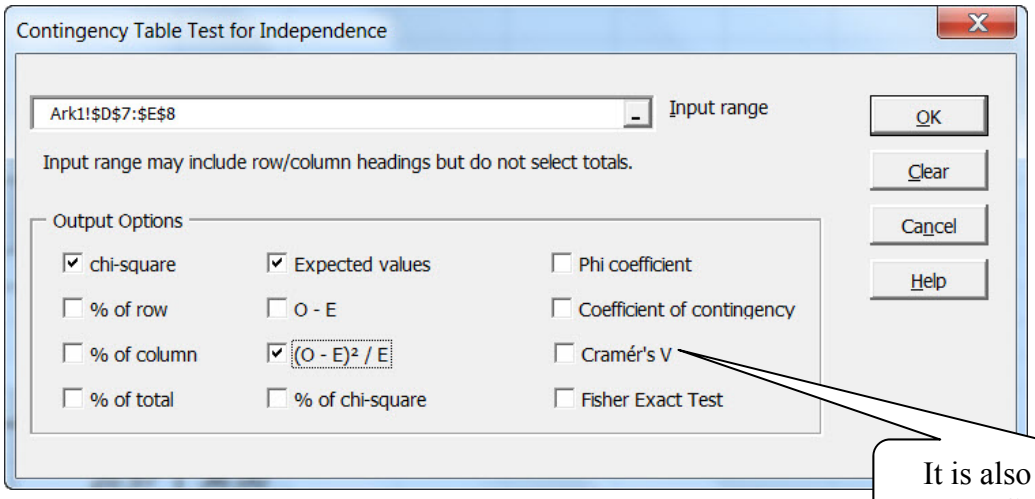
2. Mark the observed (actual) values.
3. Mark the expected values (these have manually to be calculated in advance)!
4. Click ”ok” and the p-value occurs (if it is below 0.05 we find significance at the 5 % level and accept H_1).

Performing this on our little example above gives us the following screenshot:



Chi-Square Test in Megastat

1. Here select **Add-ins | Megastat | Chi-Square Crosstab | Contingency Table** and obtain the following dialog box.
2. Mark at the relevant boxes.



Then we obtain the following output for our little test:

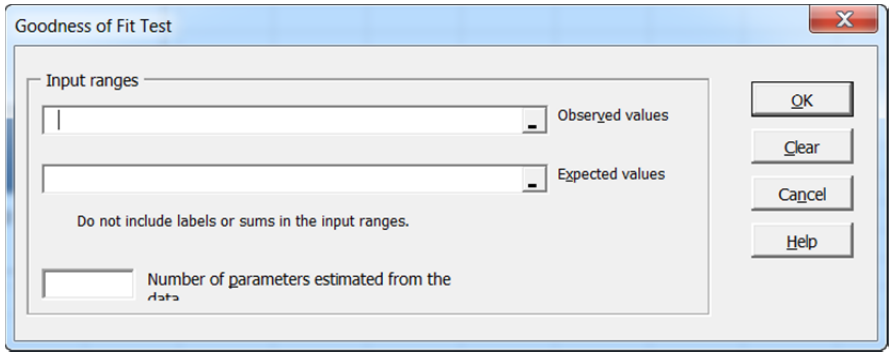
		Coke	Diet Coke	Total
Female	Observed	**3**	**33**	36
	Expected	15.43	20.57	36.00
	(O - E)² / E	10.01	7.51	17.52
Male	Observed	**30**	**11**	41
	Expected	17.57	23.43	41.00
	(O - E)² / E	8.79	6.59	15.38
Total	Observed	33	44	77
	Expected	33.00	44.00	77.00
	(O - E)² / E	18.80	14.10	32.90

32.90	chi-square
1	df
0.00	p-value
.654	Cramér's V

We can also perform a ***goodness of fit test*** in Megastat:

1. Here select **Add-ins | Megastat | Chi-Square Crosstab | Goodness of fit test.**

2. Here a dialog box very similar to the function under Excel is obtained. So we have in this case manually to calculate the expected values.



If the observed values and the expected values are loaded into the program the following output is obtained:

Goodness of Fit Test

observed	expected	O - E	(O - E)² / E	% of chisq
10	5.750	4.250	3.141	65.09
26	23.000	3.000	0.391	8.11
33	34.500	-1.500	0.065	1.35
19	23.000	-4.000	0.696	14.41
4	5.750	-1.750	0.533	11.04
92	92.000	0.000	4.826	100.00

4.83	chi-square
4	df
.3056	p-value

The results are as expected.

Chi-Squared Test Performed by use of the Pocket Calculator

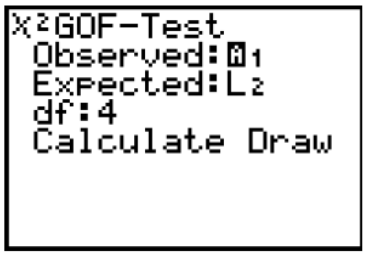
Consider first the test for *goodness of fit*[2]. Use STAT → TESTS → D: GOF-Test → ENTER. Data has to be stored in data registers for example L_1 and L_2. So the expected values have to be calculated in advance just as the case in Excel. Having stored these data then use CALCULATE → ENTER.

The output reports the chi-squared value as well as the p-value. Consider also the example below taken from the manual to the TI-84 pocket calculator:

[2] This GOF feature is not available on all TI-84 and TI-89 models. If this not is the case you can install the GOF-test by use of an update to be found on the homepage of Texas Instruments.

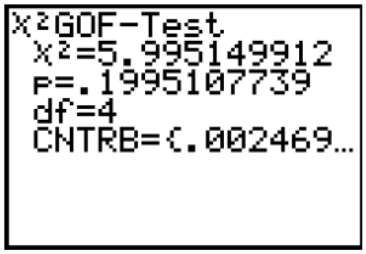
list 1={16,25,22,8,10}
list 2={16.2,21.6,16.2,14.4,12.6}

The Chi-square Goodness of Fit input screen:



Note: Press STAT ▶ ▶ to select **TESTS**. Press ▼ several times to select **D:X²GOF-Test...** Press ENTER. To enter data for df (degree of freedom), press ▼ ▼ ▼. Type 4.

Calculated results:



Next consider the ***Test for Independence***. Use STAT → TESTS → C: χ^2 -Test → ENTER

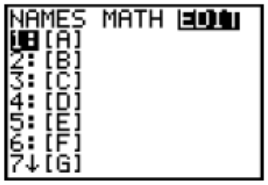
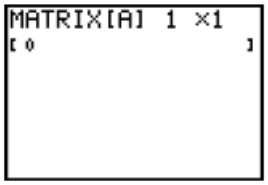
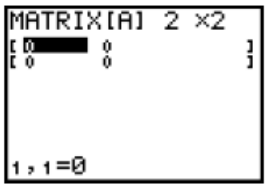
Also in this case the expected values have to be calculated in advance. As the test for independence considers the relation among two data sets, each with more than a single category data has to be displayed in matrix from. Denote the observed and expected values by matrix [A] and [B] Use then CALCULATE → ENTER.

The output displays the value for the chi-squared tester as well as the *p-value*. The number of degrees of freedom is automatically calculated.

How is data set up in matrix form? Use TEST 2ND → MATRIX. Then a menu appears where it is possible to define a matrix. Use EDIT and 1: [A] → ENTER. Initially, the dimension of the matrix has to be determined. Next he observed values are plotted in. Similarly for the expected values in the matrix [B].

Look at the following example taken from the manual to the TI-84 pocket calculator:

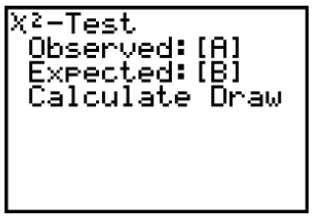
Creating a new matrix

Press	Result
2nd [MATRIX] ◂	
ENTER	
2 ENTER 2 ENTER	
1 ENTER 5 ENTER 2 ENTER 8 ENTER	

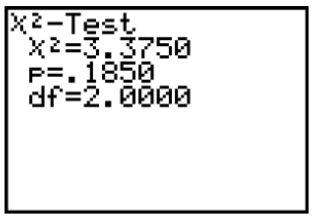
Note: When you press ENTER, the cursor automatically highlights the next cell so that you can continue entering or editing values. To enter a new value, you can start typing without pressing ENTER, but you must press ENTER to edit an existing value.

An example on how to perform the test:

Input:



Calculated results:



Nils Karl Sørensen ©
Statistics II, BA-INT

Set 10: Simple Regression Analysis

by Nils Karl Sørensen

Outline

Sections marked with an * will not be subject to independent exam questions.

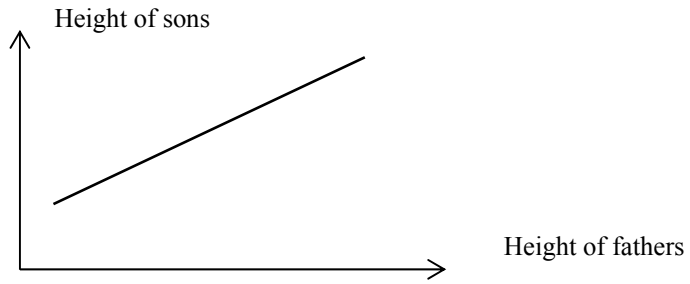
1. Setting up the Simple Regression Model and Assumptions

Regression is one of the most frequently used methods in statistics to analyze the relation among two or more variables. Regression is used to predict or forecast the outcome of a stochastic variable y by use of one or several stochastic variables x. Regression analysis make use of many of the topics considered in the previous sets of notes for example correlation, t-tests and p-values.

The regression analysis falls as a natural extension of *descriptive statistics*. In this topic, the relation among the single variables was described by use of histograms and by use of measures of location and dispersion. In addition, we considered the simple relation among two variables by use of the covariance and the correlation.

Now we are able to combine the analysis of all the variables into a statistical consistent frame of reference.

The word *regression* is diverting from the word *regress* i.e. "to return to something". The motivation for the use of the word origins from the cousin of Charles Darwin, Francis Galton. In 1889 Francis Galton published a book about heredity. In the book Francis Galton had a scatter plot with "height of fathers" measured along the horizontal x-axis, and the "height of fathers" measured along the vertical y-axis.



Galton expected a positive relation among the two variables. In addition, due to the theory of evolution he also expected a coefficient of the slope to be larger than one. This should indicate that the sons had higher height than their fathers due to better evolution etc.

To his surprise Galton found that the straight line best fitted to the points had a *slope less than one.* Tall fathers had sons that, on average, were smaller than themselves while small fathers small fathers has sons that, on average, were taller than themselves. In short, there is a going back or *regression* of heights against the middle.

The word *regression* is stuck in the statistical vocabulary. But Galton's regression is based upon a fallacy that lies close at hand when two measurings are separated in time. The natural variation in sub-groups from a population will mean that some fathers without a genetic predisposition far "tallness" are randomly tall. It is most likely that such fathers will have sons that are smaller than themselves.

On the other hand, it is most likely that small fathers without a genetic predisposition for "smallness" will have sons that are taller than themselves. It is this phenomenon that causes Galton's regression against the middle.

Setting up the Simple Regression Model

As mentioned above regression analysis is an examination among of a variable y that is "regressed" on one or more variable(s) x. In *simple regression* we consider the case of one x variable, whereas in *multiple regression* several x variables are considered. In economics the variable y is the endogenous variable whereas the variable(s) x is the exogenous variable(s).

The simple regression model with one explanatory variable *x* is more formally writes as:

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon \qquad \text{where } \varepsilon \approx N(0,\sigma^2)$$

The coefficients β_0 and β_1 are to be estimated. ε is the error term or residual. The residual is the part of the variation that not is explained by the model. The error term is the distance between the observed *y* and the expected value of *y* obtained by substitution of *x* into the regression equation, and by use of the parameters *(β´s)*. This expectation is labeled $E(y)$ or $\hat{y}$. Consequently the error term is equal to $\varepsilon = \hat{y} - y$.

In order for the estimates of the parameters not to be biased (or BLUE Best Linear Unbiased Estimates, see next page) the following assumptions for the residuals should be valid:

- The relation among *x* and *y* should be linear
- The residuals should be normally distributed, i.e. $\varepsilon_i \approx NID(0,\sigma_\varepsilon^2)$
- The residuals should be independent. The value of the error term ε corresponding to an observed value of *y* should be statically independent of the value of the error term corresponding to any other observed value of *y*.

The second bullet can be elaborated further:

- The expected value (and the mean) of the residuals should be equal to zero ("white noise")
- The variance of the residuals σ_ε^2 should be constant

For a given observation *i* the ***simple regression*** model can be stated as*:*

$$y_i = E(y_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{where } i = 1,2,\ldots,n$$

The ***multiple regression*** model with *k* "regressors" can be stated as:

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

We consider the same assumptions as with the simple regression analysis. We shall return to the multiple regression model in notes set 11. The mathematical solution to the multiple regression model is complicated a beyond the scope of this course.

2. Estimating the Simple Regression Model

In order to estimate the model we try to minimize ε_i^2. We minimize the square because then positive and negative residuals will not cancel out. So for the β*'s* we solve the problem:

$$\text{Minimize } \varepsilon^2 \text{ subject to } \beta_0 \text{ and } \beta_1 \text{ or } \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

We find the partial divertive with regard to the parameters. This is a system of two equations to be solved. The solutions are called the *normal equations.* This is complicated, so we will not do it. Instead let us solve the problem in the following way by defining the "sum of squares" for the variables *x, y* and the multiplex *xy.* Then:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{where } \bar{x} \text{ is the mean of } x$$

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{where } \bar{y} \text{ is the mean of } y$$

$$SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

The regression parameters are then found as:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \text{ and } \beta_1 = \frac{SS_{xy}}{SS_x}$$

The solution of the regression parameters by use of this method results in a minimization of the sum of the squared errors. This solution method is called *OLS* or ***O****rdinary* ***L****east* ***S****quares*[1]. A special result is

The Gauss-Markow Theorem:
The ordinary least square **(OLS)** method gives the best linear unbiased estimate **(BLUE)** of the β's

This is important in a statistical sense, because now we know for sure that our method is the optimal one given the assumptions of the regression model. Several other estimators with specific special features are used in statistics. The most common used alternative is the maximum likelihood estimator. These estimators are not a part of the course.

Having found the parameters we can find a particular value of the expected value of *y* by substitution into the regression line:

[1] This method was developed by Carl Friedrich Gauss (1777−1855) known from for example the development of the Normal distribution. He used matrix algebra to solve the system. Another mathematician Thomas Jordan developed a very elegant solution to the matrix problem. OLS is sometimes also referred to as the Gauss-Jordan method.

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

This expression can be substituted into minimization term above. We then obtain the sum of squared errors (SSE). So:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Similarly we can find the variation explained by the regression line as:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

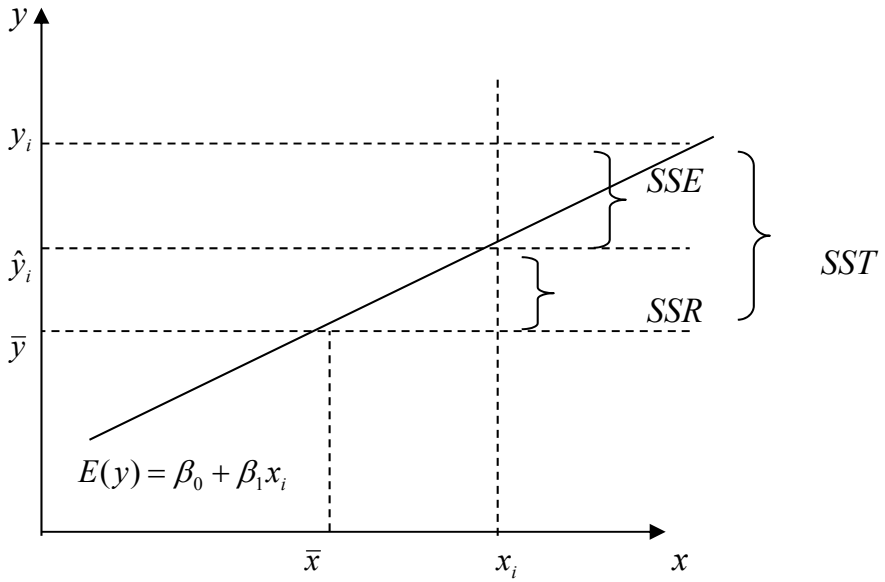
Finally the total variation is equal to:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

In general it can be shown that *SST* = *SSR* + *SSE*. This is similar to the one-way analysis of variance (ANOVA) considered in notes set 8.

Later in the present set of notes, we use this information to develop an overall F-test for regression significance. The decomposition of *SST* is shown in Figure 1.

Figure 1: Decomposition of SST



Here it is evident that *SST* = *SSR* + *SSE*.

3. Numeric Example for the Simple Regression Model

Assume the following data set for the variable y_i and x_i:

y_i	40	50	50	70	65	65	80
x_i	100	200	300	400	500	600	700

As evident when y increases so will x. So we must expect a positive regression. Further we have 7 pairs of observations.

y_i	x_i	$(x_i - \bar{x})$	$SS_x = (x_i - \bar{x})^2$	$(y_i - \bar{y})$	$SS_y = (y_i - \bar{y})^2$	$SS_{xy} = (x_i - \bar{x})(y_i - \bar{y})$
40	100	−300	90,000	−20	400	6,000
50	200	−200	40,000	−10	100	2,000
50	300	−100	10,000	−10	100	1,000
70	400	0	0	10	100	0
65	500	100	10,000	5	25	500
65	600	200	40,000	5	25	1,000
80	700	300	90,000	20	400	6,000
$\sum y_i$=420	$\sum x_i$ = 2,800		$\sum$ = 280,000		$\sum$ = 1,150	$\sum$ = 16,500

Calculating the means: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{420}{7} = 60$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2{,}800}{7} = 400$

The estimates of the coefficients can now be obtained by substitution as:

$$\beta_1 = \frac{SS_{xy}}{SS_x} = \frac{16{,}500}{280{,}000} = 0.059$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 60 - 0.059(400) = 36.4$$

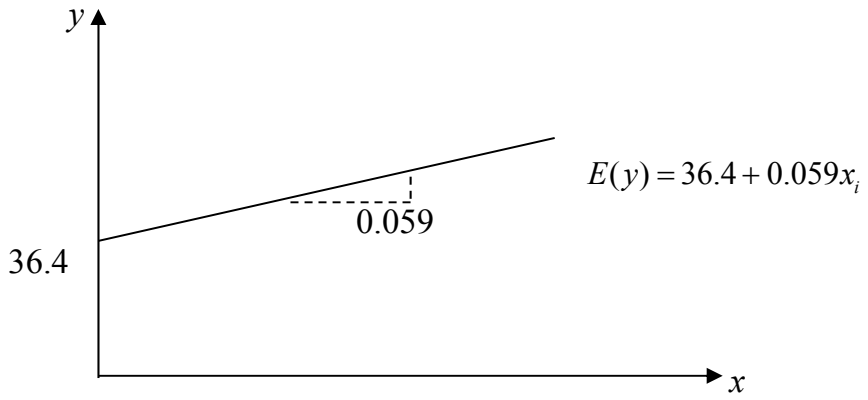
So the solution is equal to: $\hat{y}_i = \beta_0 + \beta_1 x_i = 36.4 + 0.059 x_i$

This is a very flat curve with a positive slope. The illustration is provided in the next page.

We shall continue to use this example for the remaining parts of these notes.

The simple (as well as the multiple) regression model can be estimated by use of Excel, Megastat or by use of the Pocket calculator, see Section 9 of the present set of notes.

Figure 2: Illustration of Example



4. The Coefficient of Correlation

For the simple regression model several interesting measures can be calculated. In this section we consider the coefficient of correlation, the coefficient of determination and the standard error.

The Coefficient of Correlation

First, the *coefficient of correlation* as known from the note set 2 on "Correlation and Covariance". The coefficient is also called multiple R in the Excel output. As remembered from notes set 2 the coefficient of correlation is defined as:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

In this way $-1 \le r \le 1$. If correlation is negative then a plot between x and y will have a negative slope and vice versa. If the coefficient of correlation is zero then the variables are not related.

We can calculate the correlation in our example as:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{16{,}500}{\sqrt{280{,}000(1{,}150)}} = 0.9195$$

This is positive as expected.

We can perform a test of the significance of the correlation or population correlation coefficient. This test can be used to determine the overall significance of the regression. Let us denote the population correlation coefficient by symbol ρ (rho). The two-sided test of whether two random variables x and y are correlated can then be formulated as:

$H_0: \rho = 0$	no relation among the variables
$H_1: \rho \neq 0$	positive or negative relation among the variables

The tester has a bivariate normal distribution approximated by a t-distribution with degrees of freedom equal to $n-2$. The tester can be written as:

$$t_{(n-2)} = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

In our example:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.9195}{\sqrt{(1-0.9195^2)/(7-2)}} = \frac{0.9195}{0.1758} = 5.23$$

At $\alpha = 0.05/2 = 0.025$ and $df = 7-2$ then $t_{0.025(5)} = 2.571$. So $t > t_{0.025(5)}$ $(5.23 > 2.571)$ implying that H_1 *is accepted* and the relation is significant. With the high correlation this is expected.

The Coefficient of Determination

First, the *coefficient of determination* was also introduced in note set 2 on "Correlation and Covariance". The *coefficient of determination* (or R square) is defined as:

$$R^2 : 1 - \frac{SSE}{SST} \qquad \text{or} \qquad R^2 = r^2$$

It is just the square of the coefficient of correlation. The range of R^2 is $0 \leq R^2 \leq 1$. If R^2 is low then the model is poor and vice versa.

Again we can apply by use of the example. First:

$$R^2 = r^2 = (0.9195)^2 = 0.8455$$

The interpretation is that the in the model x explains 84.55 percent of the variation in y.

In order to calculate the first formula we need to calculate *SSE* and *SST*. This is a more complicated. We can modify our formulas as follows:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = SS_y - \frac{SS_{xy}^2}{SS_x} = SS_y - \beta_1 SS_{xy}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_y$$

Then:

$$SSE = SS_y - \beta_1 SS_{xy} = 1{,}150 - (0.059)16{,}500 = 177.68$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_y = 1{,}150$$

By substitution:

$$R^2 : 1 - \frac{SSE}{SST} = 1 - \frac{177.68}{1150} = 0.8455$$

The Adjusted Coefficient of Determination

If the model is expanded by inclusion of many variables this measure is not optimal. This is the case in multiple regression. To correct for this "inclusion of many *x* variables" on R^2 we introduce the *adjusted* R^2 defined as:

$$R_{adj}^2 : 1 - \frac{SSE/[n - (k+1)]}{SST/(n-1)}$$

Here *k* is the number of included regressors (number of *x* variables), and *n* is the number of observations. The adjusted coefficient of correlation is smaller than the "unadjusted".

In our example *k = 1* (the number of explanatory *x* variables). By substitution:

$$R_{adj}^2 : 1 - \frac{SSE/[n - (k+1)]}{SST/(n-1)} = 1 - \frac{177.68/[7 - (1+1)]}{1150/(7-1)} = 1 - \frac{35.536}{191.667} = 0.8146$$

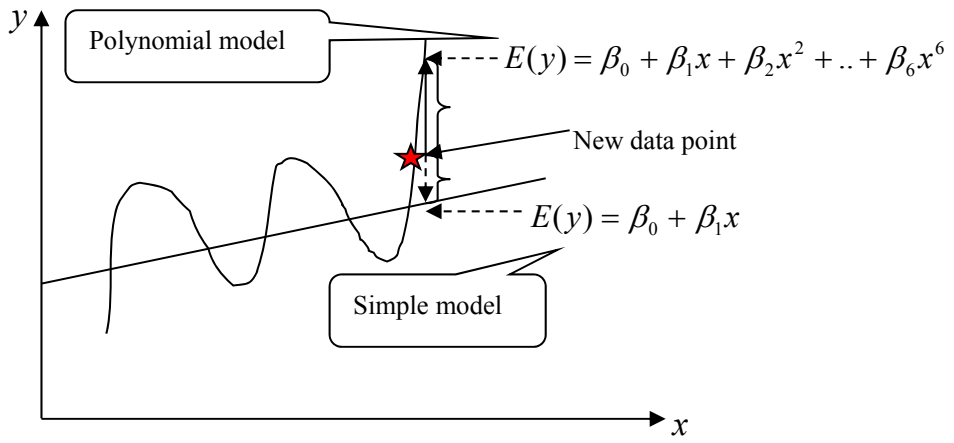
This is smaller than R^2 and also what we expected. Notice that this correction will be smaller when the number of observations increases.

Discussion of the Adjusted Coefficient of Determination*

What happens when the number of independent or explanatory variables is increased too much? In such a case we are confronted with a situation of *overfitting*. Let us consider our numeric example above in Section 3. However, now instead of one independent variable we have six (*k=6)* independent variables (*x*-variables), and still *n* = 7 observations. As each variable adds extra information in order to explain *y* we must expect R^2 to take a high value (maybe 0.999). Let us now inspect the degrees of freedom in this case. In the simple linear regression (k=1) we have degrees of freedom equal to *(n − (k+1)) = (7 − 2) = 5*. This is so

because we have one independent variable *x* plus the constant term. Consequently in the new case, we have *(7 − (6+1)) = 0* degrees left for modeling the error! This implies that the values of the parameters are extremely uncertain, and predictions cannot be performed.

Figure 3: Overfitting a Data Set with a 6-Degree Polynomial



The idea of overfitting can be seen in another way. Multiple regression includes a technique called polynomial regression. In polynomial regression, we regress a dependent variable *y* on powers of the independent variables. Each power of a variable, say, x_1^2, is considered as an independent variable in its own right. For example, a six-degree polynomial regression in *x* is modeled as $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \varepsilon$. This is treated as a six-variable regression equation.

Overfitting in this context is shown in Figure 3. As shown, a 6-degree polynomial provides a perfect fit to a set of 7 data points (notice *df = 0*). The prediction of a new data point using this model is worse than the prediction provided by a 1-degree model, i.e. a least square line estimated in the example in Section 3. The line captures the systematic trend in the data shown in Figure 2 leaving the rest of the movements within the data to error. In this case, the straight line is a good statistical model, whereas the 6-degree polynomial is not.

Consider the case of the prediction of a new data point indicated by the point marked by the solid arrow and the star in the diagram. The simple model results in a smaller prediction error shown by the dotted arrow, than the error resulting from the polynomial model. Overall, the volatility of the predictions performed by the polynomial model is much larger than the volatility of the predictions performed by the simple model.

The Standard Error

The standard error is an important measure of performance, and is used when models are compared. This measure captures the problems with the overfitting present when using the coefficient of determination. The standard error in the regression defined as:

$$s = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}} = \sqrt{MSE}$$

For definition of the mean square error *MSE,* see next section page 12. The standard error is an important measure because it is the normalized number that we want to minimize. *When inspecting an output from a regression we should rather give attention to the standard error than to the determinant of determination.*

In our example:

$$s = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{177.68}{7-(1+1)}} = 5.96$$

This is by the use of the definitions given for the multiple regression case easily expanded to this case.

What is found for a Good Regression Model?

Summing up for a regression model to be good the following conditions should be met:

- The correlation should be high
- The determination should be high
- The standard error should be as small as possible

In general the last condition is the most important.

The relative value of for example R^2 may vary with the size of the data set. For example in a data set with 2,000 observations much more variation may be present than in a data set with 20 observations. Therefore even a good model with significant parameters will may a lower R^2 than the R^2 obtained from the regression of the same model with 20 observations and less significant parameters.

5. Testing for Significance

We can set up two tests in order to investigate the model namely the ***F-test*** and the ***t-test.*** A test is set up of three elements:

1. Setting up hypothesis
2. Performing the test
3. Evaluation of the test

The F-test for Overall Significance

This is a general test for the full model, i.e. a relationship between y and any of the x_i. Define the hypothesis:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (the regression has no meaning)
H_1: At least one of $\beta_1, \beta_2, \ldots, \beta_k$ is different

The F-test is conducted from the ANOVA analysis outlined in note set 8. It involves a decomposition of the variation of the data into a part explained by the regression and the non-explained part namely the error. By constructing the ANOVA table as we did in the notes set 8, we can calculate the tester. Consider the table:

ANOVA-table:

Variation	*Squared sum (SS)*	*Degrees of freedom (df)*	*Mean square (MS)*	*F-value*
Regression	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	MSR = SSR/k	$F = \frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n − (k+1)	MSE = SSE/ n−(k+1)	
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	n − 1		

Notice the relation to Figure 1 shown above.

The tester is F-distributed with the degrees of freedom given in the table. We need to find the critical value of F-distribution from **Statistics Tables**. With the data from our example we obtain:

ANOVA-table:

Variation	*Squared sum (SS)*	*Degrees of freedom (df)*	*Mean square (MS)*	*F-value*
Regression	$SSR = 972.32$	1	MSR = 972.32	$F = \frac{972.32}{35.536} = 27.36$
Error	$SSE = 177.68$	5	MSE = 35.536	
Total	$SST = 1{,}150$	6		

Our tester is *F(1,5)*. Assuming α = 0.05 then F_{crit} = 6.61 as found in **Statistics Tables.** As 27.36 > 6.61 we accept H_1. So the regression is significant. Remember that the exact *p-value* can be found as shown in the notes set 6.

The t-test for Coefficient Significance

A regression model is not likely to be useful unless there is a significant relationship among *y* and the *x* variable. Consider a partial test for the significance of a specific variable x_i. This is equivalent to examine the coefficient β_i for significance. This is properly the most useful test in regression statistics! Set up the hypothesis for a two-sided test as:

H_0: $\beta_i = 0$	the coefficient is not important
H_1: $\beta_i \neq 0$	the coefficient is important

The tester can be computed as:

$$t = \frac{\beta_i}{s(\beta_i)}$$ with degrees of freedom equal to $df = n - (k+1)$

So in the simple regression with only one explanatory variable $df = n - 2$

Further, $s(\beta_i)$ is the standard deviation of the estimate. The standard deviation is given as:

$$s(\beta_1) = \frac{s}{\sqrt{SS_x}}$$

The significance of the test can be seen directly from the output where the p-value is given. This can be given the following interpretation (remember the notes set 6):

- *Weak significance:* $p < 0.10$ (significance at 10 % level or α = 0.10)
- *Significance:* $p < 0.05$ (significance at 5 % level or α = 0.05)
- *Strong significance* $p < 0.01$ (significance at 1 % level or α = 0.01)

We can set up a confidence interval for the estimated parameter(s). This may also be used in tests. A 100(1−α) percent confidence interval for β_i is given by:

$$\left[\beta_i \pm t_{\alpha/2} s(\beta_i)\right]$$

Where α is the confidence level. For a 95 % confidence interval α = 0.05. The degrees of freedom is equal to $df = n - (k+1)$.

Finally, we set up a specific test for a given value of the parameter. For example in a consumption function we can test the hypothesis than the marginal propensity to consume is equal to a value b_i. A partial test for each coefficient β_i can then be stated as:

H_0: $\beta_i = b_i$	the coefficient is equal to b_i
H_1: $\beta_i \neq b_i$	the coefficient is not equal to b_i

The tester can be computed as:

$$t = \frac{b_i - \beta_i}{s(\beta_i)}$$ with degrees of freedom equal to $df = n - (k+1)$

Evaluation of the tester as described above.

In our example from the simple regression, we can calculate the tester. First, we have to calculate the standard deviation as:

$$s(\beta_1) = \frac{s}{\sqrt{SS_x}} = \frac{5.96}{\sqrt{280{,}000}} = 0.01126$$

The t-statistic is:

$$t = \frac{\beta_1}{s(\beta_1)} = \frac{0.05892}{0.01126} = 5.2327$$

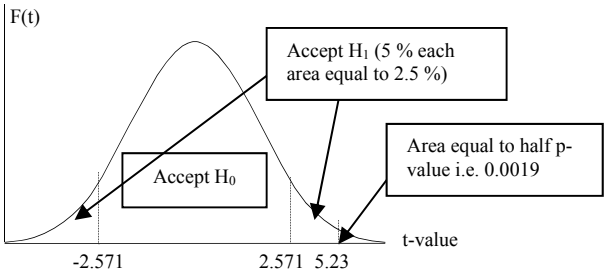
Assuming that $\alpha = 0.05$, a two-sided test and with degrees of freedom equal to $df = n - (k+1) = 7 - (1+1)) = 5$ then $\alpha = 0.05/2 = 0.025$. The critical value is then $t_{0.025(5)} = 2.571$. We then *accept* H_1.

In this special case with only a single x value the outcome of this test is similar to the test for significance of the correlation performed above.

With a t-value equal to 5.2327 we can approximate the *p-value* by use of the critical values in **Statistics Tables**. We have $df = 5$. At the row in the table at $df=5$ we find the t-value nearest to 5.2327. At $\alpha = 0.001$ we find a value equal to 5.893 and at $\alpha = 0.005$ we find t equal to 4.023. By interpolation between these values we find a critical value for α to equal 0.0019. With a two-sided test we find the *p-value* to equal $2 \times 0.0019 = 0.0038$. As this value is lower than 0.01 we observe a strong positive significance.

Due to the complexity of the t-distribution (one for each degree of freedom) we never have to calculate the p-value in regression analysis. However, to understand the concept of the p-value is very important, because it is an easy to use tool in order to determine the degree of significance of a given independent variable.

Figure 4: Illustration of t-test



We can set up a test for a specific value. If we for example has a prior that the value of the slope could equal 0.07 based on some economic theory could we then confirm or reject this value? Our hypothesis is then:

$$H_0: \beta_1 = 0.07$$
$$H_1: \beta_1 \neq 0.07$$

Notice that we in this case consider a two-sided test. However, we could as an alternative have performed a one-sided test if the theory has been stated more specific. The tester can be computed as:

$$t = \frac{b_1 - \beta_1}{s(\beta_1)} = \frac{0.07 - 0.059}{0.01126} = 0.9769$$

With *df = 5* we have a two-sided critical value as earlier equal to 2.571. We consequently *accept* H_0. This means that the statement is true at the given values, and we can accept theory. If we have considered a one-sided hypothesis the t-value had equaled $t_{0.05(5)} = 2.015$. Still we would have accepted H_0.

With regard to the two-sided test we could have as an easy alternative have inspected the 95 % confidence interval. This can be calculated as:

$$\left[\beta_1 \pm t_{\alpha/2} s(\beta_1)\right] \Rightarrow \left[0.059 \pm 2.571(0.01126)\right] \Rightarrow \left[0.059 \pm 0.0289\right] \Rightarrow \left[0.0299; 0.0879\right]$$

As the value 0.07 is within this range we accept H_0. The 95 % confidence interval is by default provided by the Excel as well as the Megastat output.

6. Confidence Interval and Prediction Intervals

A *prediction* is the expected value of the dependent value y calculated by use of the regression based on new given values of the independent value x_0. An outcome of a prediction falls normally within the range of the dependent variable considered. An *out of sample prediction* is normally referred to as a *forecast.* This is especially used when considering time series statistics.

The prediction is: $\hat{y} = \beta_0 + \beta_1 x_0$

Only by luck we will have a situation where the prediction is equal to the observed value. However, if the assumptions for the regression model are fulfilled then the population of all possible values of $\hat{y}$ is normally distributed with mean $\mu_{y|x_0}$ in the simple case and standard deviation given by:

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}$$

Where s is the standard error. The term in the square root is frequently referred to as the *distance value.* It measures the distance between the value x_0 and $\bar{x}$ the average of the previous observed values of x.

A 100(1−α) percent *confidence interval* for the predicted value of y is:

$$\left[\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}\right]$$

Where the degrees of freedom is equal to $df = n - (k+1)$.

For the example we can calculate a 95 % confidence interval assuming that for example $x_0 = 650$ as.

First we compute the prediction, and then we set up the interval:

Prediction: $\hat{y} = \beta_0 + \beta_1 x_0 = 36.4 + 0.059(650) = 74.75$

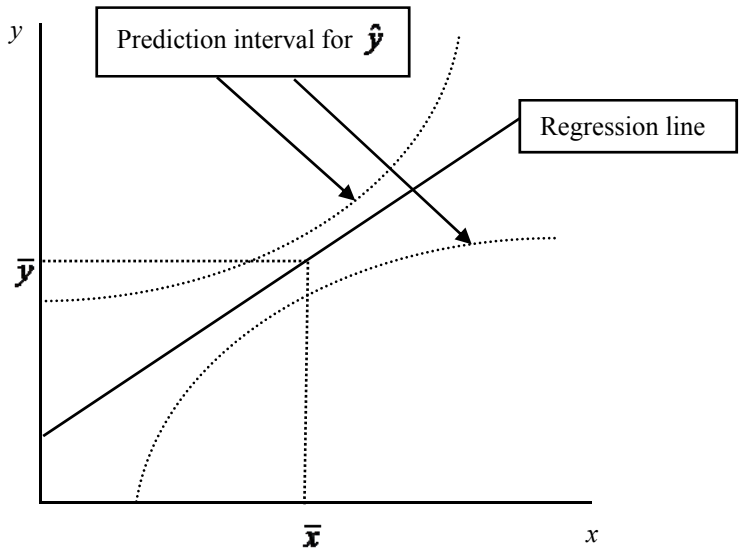
The confidence interval is found as:

$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}\right] \Rightarrow \left[74.75 \pm 2.571(5.96)\sqrt{\frac{1}{7} + \frac{(650 - 400)^2}{280{,}000}}\right] \Rightarrow \left[74.75 \pm 2.571(5.96)\sqrt{0.37}\right]$$
$$\Rightarrow [74.75 \pm 9.27] \Rightarrow [65.48\,;\,84.02]$$

Where the mean and other values are taken from the table in Section 3, further with α = 0.05 and df = 5 the t-statistic as above is equal to 2.571.

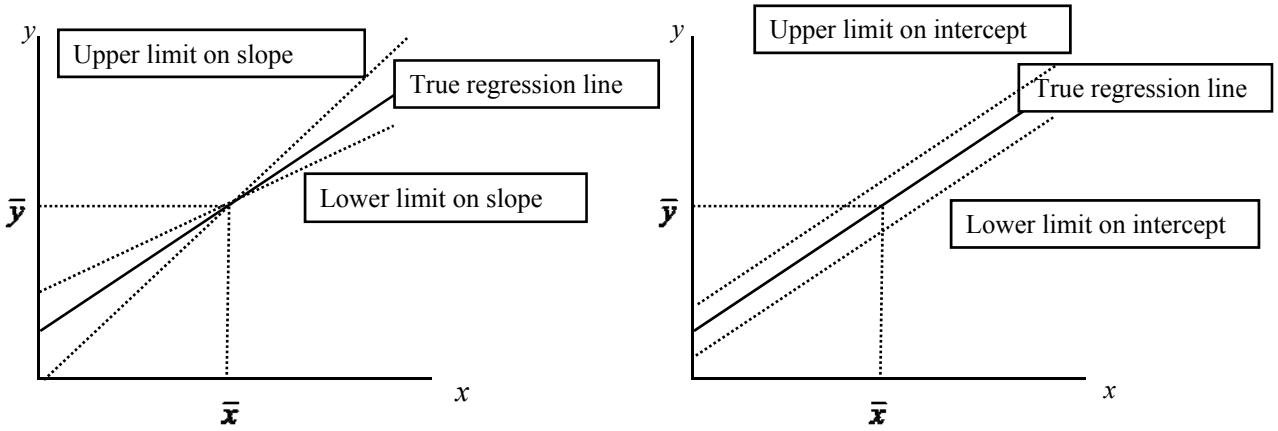
The prediction interval is illustrated in figure 5. Inspection reveals that as the distance from the mean $\bar{x}$ increases, so will the distance value increase. This implies that the size of the confidence interval will increase.

Figure 5: Illustration of Prediction Interval



In order for the prediction to be efficient it must be based on a true model. If the model is based on a sample that is biased or incomplete the estimated regression line will be wrong either with regard to the constant or the slope. In addition, the model may be non-linear of nature. Figure 5 explores what happens if either the constant or the slope is wrongly estimated. In both cases, the predictions are systematically wrong. Besides, from the two cases considered we have estimated a linear relation on a true non-linear relation. The data used may be non-linear outside the range of the sample used for the estimation.

Figure 6: Errors in Prediction



Uncertainty about the slope of the regression line

Uncertainty about the intercept of the regression line

7. Model Control and Heteroscedasticity

The model control is a *very* important part of the regression analysis. The model control consists of investigation of the residual plot (there may be more than one in multiple regression), and considering the plot for normal distribution. The residual plots should exhibit "white noise", whereas the normality plot should show a cumulated straight line.

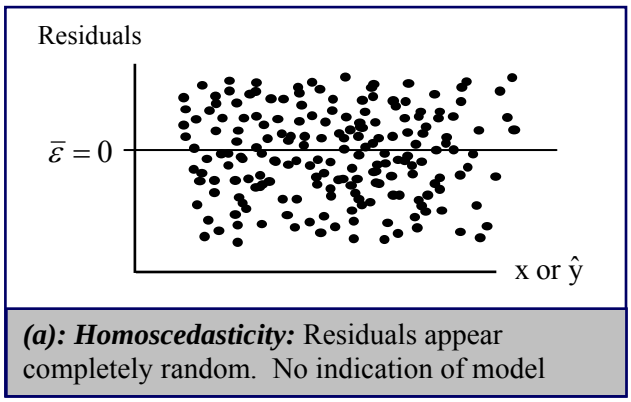
The goal of the model control is to check that the assumptions with regard to the error term is fulfilled, i.e. mean equal to zero, constant variance and normally distributed errors.

Figure 7 gives some examples of residual plots. First, the ideal case is the *homoscedastic* plot in *panel a*. Residuals are grouped as "white noise" around a zero mean with constant variance. The following panels show some typical errors. If we observe patterns of this kind our estimates the estimates of the parameters are biased and consequently not BLUE.

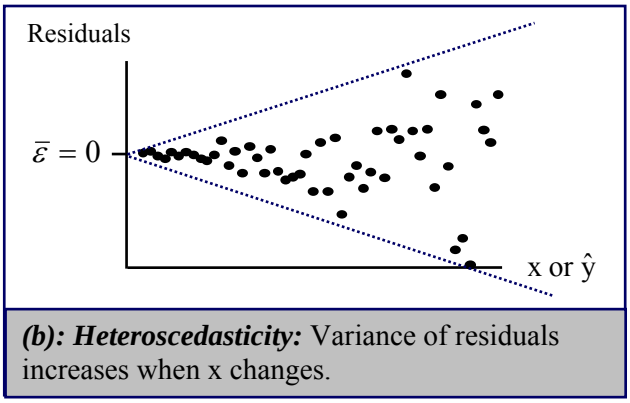
- *Panel (b):* The residuals are not constant in variation and the magnitude increases or decreases relative to the horizontal axis. The assumption of constant variance of the residuals is violated. This is called *heteroscedasticity*.
- *Panel (c):* The mean is not stationary. There is typically dependence from one observation to the next. This is very likely to occur in time series data.
- *Panel (d):* The underlying data are non-linear. We can correct for this in various ways for example by taking the logarithm to *x* and *y* and provide a *log-linear* estimation.

The final assumption to inspect is the one for normality. Software such Excel and Megastat undertake this issue by computing a plot for normality. On the horizontal axis we find the cumulative residual values and z-values, whereas Excel plots the corresponding values of *y* on the vertical axis. If the residuals are normal then they should align themselves along a straight line cumulative in the plot.

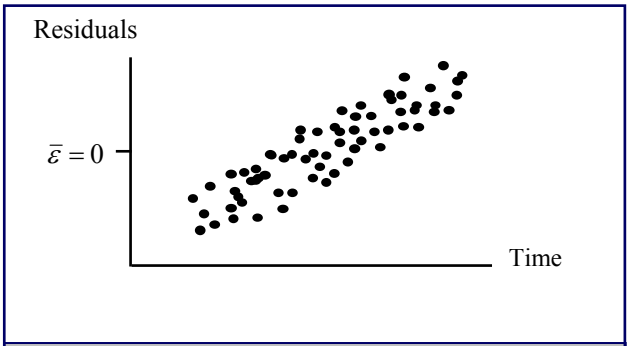
Figure 7: Examples of Residual Plots



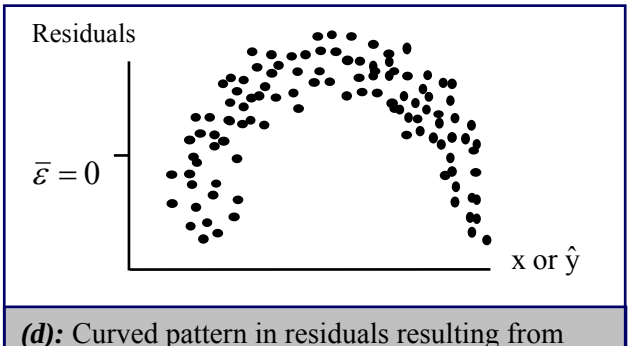
(a): Homoscedasticity: Residuals appear completely random. No indication of model



(b): Heteroscedasticity: Variance of residuals increases when x changes.



(c): Residuals exhibit a linear trend with time.



(d): Curved pattern in residuals resulting from underlying non-linear relationship.

Construction of the Normal Probability Plot*

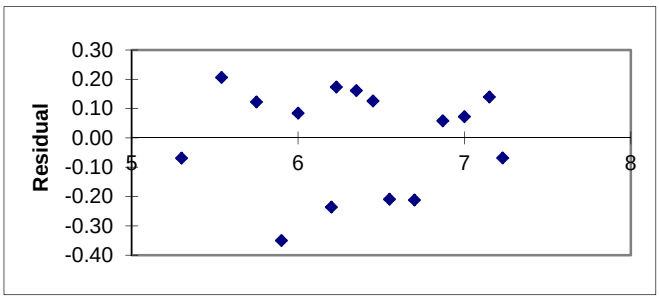
The *normal probability plot* is constructed as follows. For each value ε of the residual, its quartile (i.e. cumulative probability) is calculated using the equation:

$$q = \frac{l + 1 + m/2}{n + 1}$$

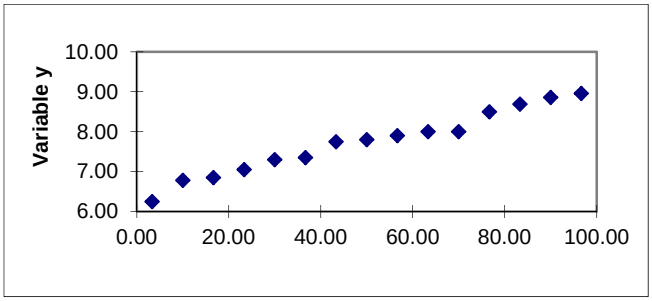
where l is the number of observations less than ε, m is the number of residuals equal to ε, and n is the total number of observations. Then the z value corresponding to the quartile denoted z_q is calculated. Typically then z_q and y is calculated. This depends on the computer program/template. Some programs plot z_q versus ε. The process is repeated for all residuals and the result is plotted in a diagram.

With regard to this plot, it is sufficient to know that the diagram should be a relatively straight cumulated line. Relative to the output from Excel data are being cumulated and ranging from 0 to 100.

Let us look at an example of residual plot and normality from a simple regression:



The residuals are not perfect, but sufficient. Notice that there may be some signs of heteroscedasticity. For normality a reasonable plot could look like the one below:



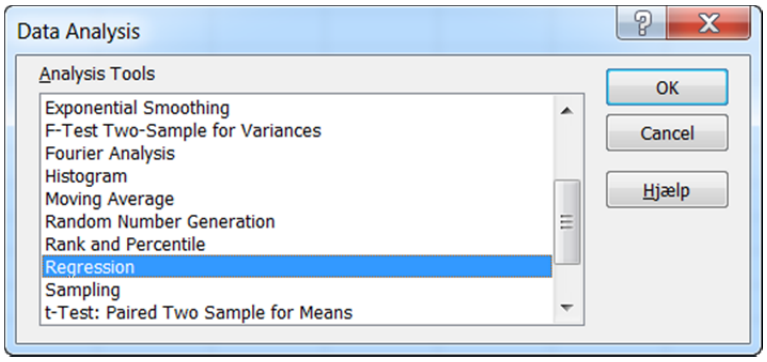
8. Regression using Excel, Megastat or the Pocket Calculator

The simple regression analysis can be performed in Excel, Megastat and also by use of the Pocket calculator.

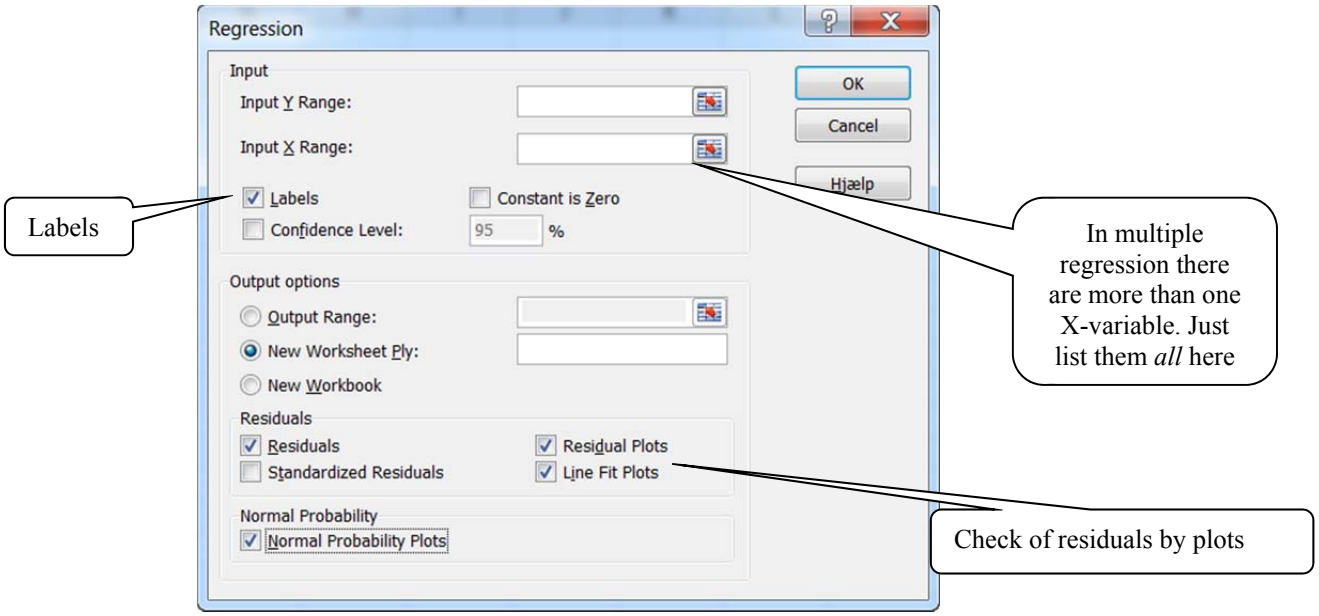
Excel

In Excel we can perform a regression by the following sequence:

1. Select **data | data analysis | regression.** The following menu appears:

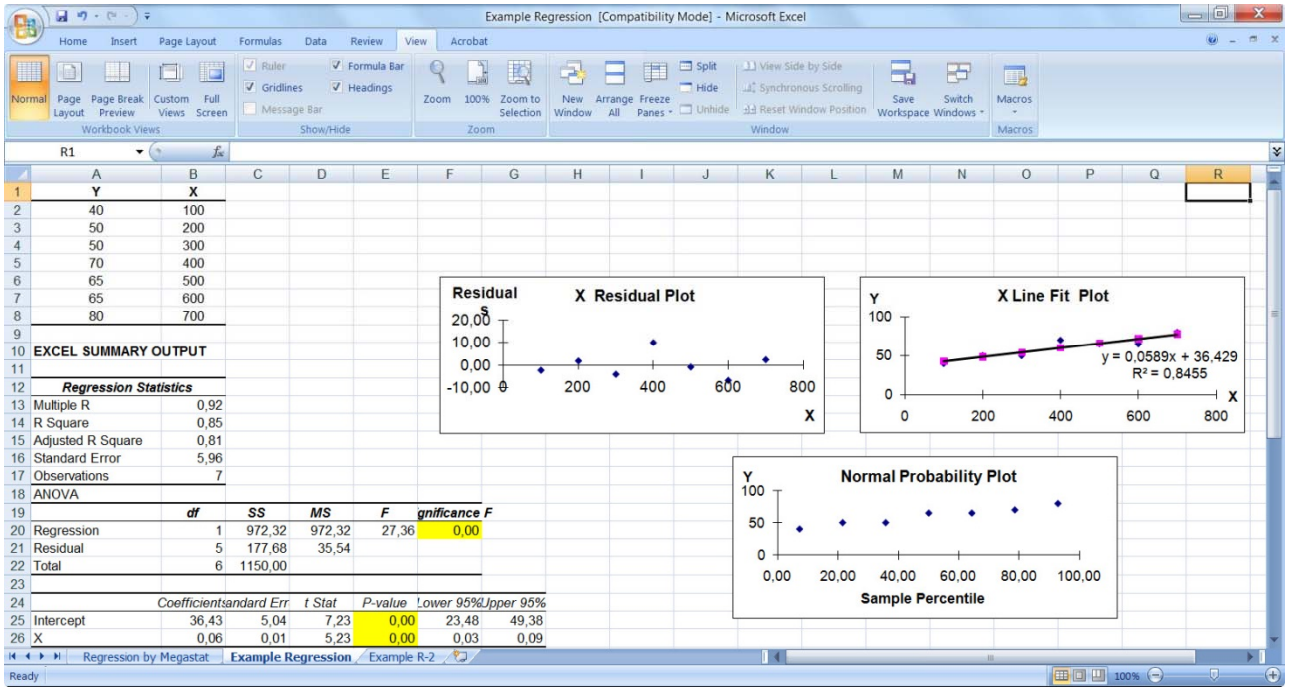


Obtain the box:



2. Select input range for Y and X. If **multiple regression** is performed *all* X-variables are loaded into the area X.
3. Mark ”labels” and ”residuals”, ”residual plots” and ”normal probability plots” if you want to check the model.
4. Press ”ok”.

Then we obtain the following screenshot (the yellow marks are my own):



	Y	X
2	40	100
3	50	200
4	50	300
5	70	400
6	65	500
7	65	600
8	80	700

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,92
R Square	0,85
Adjusted R Square	0,81
Standard Error	5,96
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	972,32	972,32	27,36	0,00
Residual	5	177,68	35,54		
Total	6	1150,00			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	36,43	5,04	7,23	0,00	23,48	49,38
X	0,06	0,01	5,23	0,00	0,03	0,09

Regression output looks like (except for the yellow signature):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9195
R Square	0.8455
Adjusted R Square	0.8146
Standard Error	5.96
Observations	7

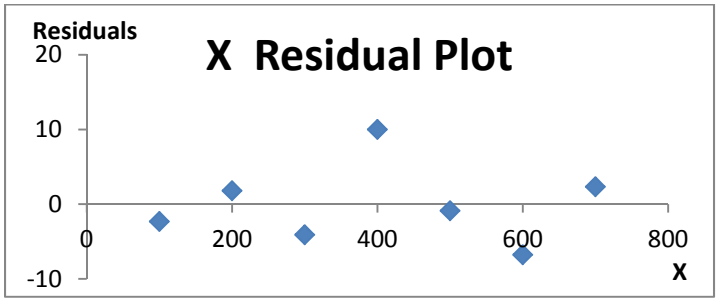
ANOVA

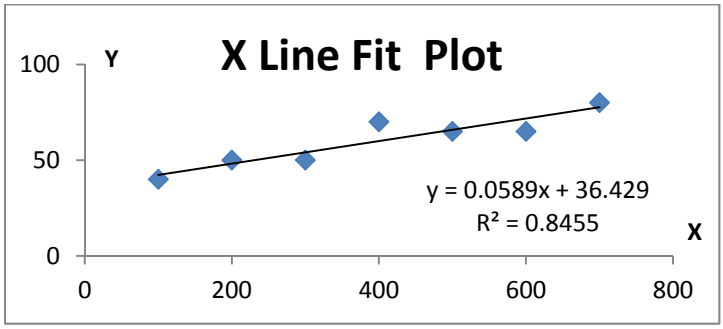
	df	*SS*	*MS*	*F*	*Signifi F*
Regression	1	972.32	972.32	27.36	0.00
Residual	5	177.68	35.54		
Total	6	1150			

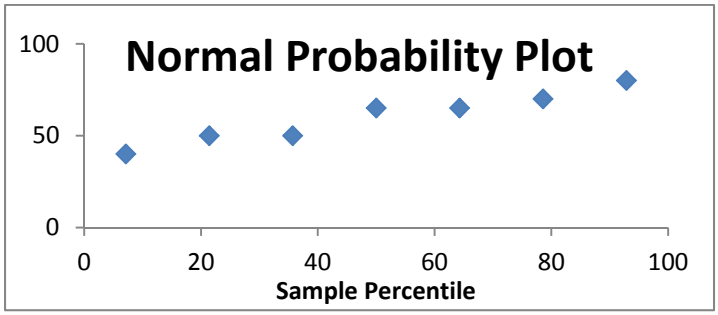
	Coef	*St Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	36.43	5.04	7.23	0.00	23.48	49.38
X	0.0589	0.0113	5.23	0.00	0.0300	0.0879

Observe that the numbers that has been calculated from the example in the text is found here. The p-values are all "0.00" (I have formatted the numbers to two digits). As they are below 0.01 strong significance is found. Just as expected.

The plots of residuals, line plot and plot for normality is found as displayed. Below. I have undertaken some editing, and for the line plot I have used the function "add trendline" and have inserted the formula and the R^2 value.





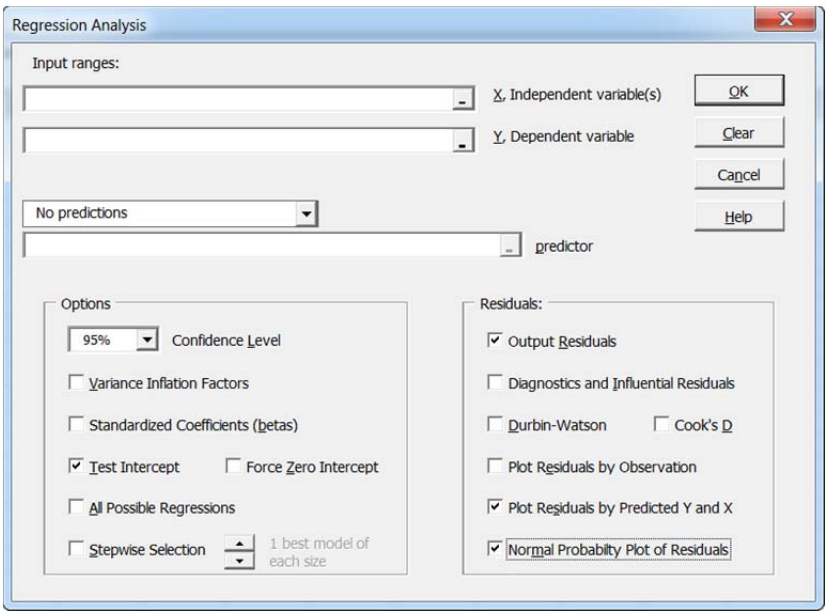


The plots look to be in order ☺

Megastat

In Megastat we can perform a regression by the following sequence:

1. Select **Correlation / regression | regression analysis.** The following menu appears:



2. Select input range for X and Y (Notice that it is reverse relative to Excel). If **multiple regression** is performed *all* X-variables are loaded into the area X.
3. Mark ”output residuals”, ”plot residuals” and ”normal probability plots” if you want to check the model.
4. Press ”ok”.

The output look as:

Regression Analysis by Megastat

r^2	0.845	n	7
r	0.920	k	1
Std. Error	5.961	Dep. Var.	**Y**

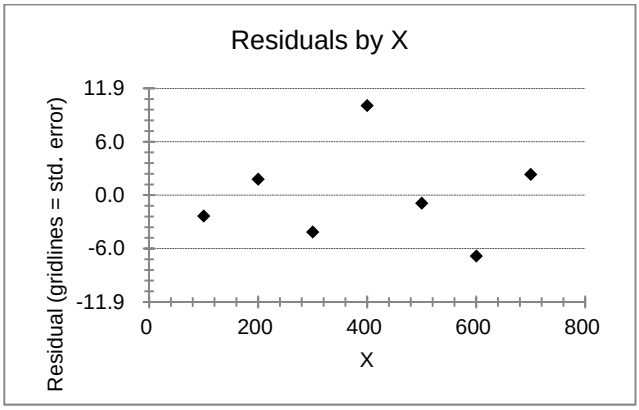
ANOVA table

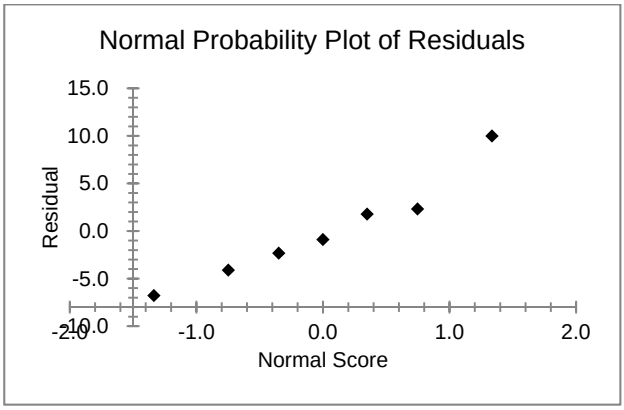
Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	972.3214	1	972.3214	27.36	.0034
Residual	177.6786	5	35.5357		
Total	1,150.0000	6			

Regression output

					confidence interval	
variables	*coefficients*	*std. error*	*t (df=5)*	*p-value*	*95% lower*	*95% upper*
Intercept	36.4286	5.0381	7.231	.0008	23.4777	49.3795
X	0.0589	0.0113	5.231	.0034	0.0300	0.0879

Residuals and plot for normality:





Pocket Calculator

On the **Pocket Calculator** only simple regression analysis can be performed. For example load data for the variable *x* into the register *L1*, whereas *y* could be in the registerkan *L2*. Now it is possible to proceed along two roads:

I. Tast STAT → CALC → 4: Linreg(ax+b) → ENTER. Now the format is Linreg(ax+b) *L1,L2* → ENTER

The result is a screen output with the values of the coefficients (labelled a and b), R^2 (coefficient of determination) and r (correlation).

II. Tast STAT → TESTS → F: LinregTTest → ENTER. Now there is a screen picture where you have to state Xlist: *L1* and Ylist: *L2*. Go to CALCULATE → ENTER

Here a more efficient output is performed with the degrees of freedom, t-tester, p-value, value of coefficients, standard error, R^2 (coefficient of determination) and r (coefficient of correlation).

In both cases it is important to remember in which register x and y is located!

Nils Karl Sørensen ©
Statistics II, BA-INT

Set 11: Multiple Regression Analysis

by Nils Karl Sørensen

Outline

Sections marked with an * will not be subject to independent exam questions.

1. Introduction

Multiple regression, is the extension of simple regression, to take account of more than one independent variable *x*. It is obviously the appropriate technique when we want to investigate the effects on *y* of several variables simultaneously. Yet, even if we are interested in the effect of only one variable, it usually is wise to include other variables influencing *y* in a multiple regression analysis for two reasons:

- To reduce stochastic error, and hence reduce the residual variance.
- Even more important, to eliminate bias that might result if we just ignored a variable that substantially affects *y*.

The multiple regression model with *k* regressors can in a mathematical way be stated as:

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

We consider the same assumptions as with the simple regression analysis with regard to the error term. As stressed above it is difficult to solve for the *β's*. This is so because we now consider a system of *(k+1)* normal equations.

The multiple regression is estimated similar to the simple regression model. The only change is that now all *x* variables are loaded into the box for "x-variable input".

With multiple regression we can perform more detailed analyses of the variables.

2. Example of Multiple Regression Output

Statistics collected at the individual level is increasingly being used in economic and social analyses; for example when the effect is examined of education and job experience on the income level. Especially many American analyses examine the influence of color and belonging to the Southern States.

In order to analyze such data sets multiple regression is frequently being used. Consider as an example a data set for 359 US citizens with the following variables:

Variable	***Description***
INCOME	Annual net income in USD
EDU	Number of years with formal education (including primary school)
EXP	Years of work experience
YEAR	Number of years at the present employer
IQ	Intelligence Quotient Index
MARRIED	Variable is 1 if the person is married otherwise 0
SOUTH	Variable is 1 if the person is from the South otherwise 0
CITY	Variable is 1 if the person is from a city area otherwise 0
COLOR	Variable is 1 if the person is colored otherwise 0

The four last variables take only the values 0 or 1. This is called a *dummy variable.*

Let us look at a multiple regression of an income model. We regress *INCOME* on the variables *EDU, EXP, YEAR, IQ, MARRIED, SOUTH, CITY* and *COLOR* in order to examine the impact.

The result is the following output:

Regression Statistics	
Multiple R	0.47
R Square	0.22
Adjusted R Square	0.20
Standard Error	6228.29
Observations	359

ANOVA

	df	*SS*	*MS*	*F*	*Signif F*
Regression	8	3.84E+09	4.8E+08	12.37	0.00
Residual	350	1.36E+10	38791613		
Total	358	1.74E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-12049.66	3411.32	-3.53	0.00	-18758.93	-5340.39
Edu	1015.05	189.95	5.34	0.00	641.48	1388.63
Exp	332.00	89.41	3.71	0.00	156.14	507.85
Year	106.01	67.76	1.56	0.12	-27.26	239.28
IQ	35.53	26.69	1.33	0.18	-16.96	88.03
Married	3203.78	1027.70	3.12	0.00	1182.54	5225.01
South	842.58	701.45	1.20	0.23	-537.01	2222.17
City	1959.56	722.76	2.71	0.01	538.07	3381.05
Color	-3554.36	1106.01	-3.21	0.00	-5729.63	-1379.10

The interesting here is the interpretation of the coefficients relative to what is expected from theory.

The variables *EDU* and *EXP* are both positively significant. So longer education and more experience on the labor market increases income. This is meaningful. If a person has used more time on education the period to earn an income and save for retirement is shorter.

The variables *YEAR* and *IQ* have no influence because the p-values are above 0.10. The variable *YEAR* is related to how long a person has been at given employer. This should not have influence. *IQ* is difficult to measure, and may also be correlated to for example *EDU.* When two *x*-variable are highly correlated it is called *multicollinarity.* This feature is not good because the two variables are so interested in explain each other that they not can explain *y*. One of the variables should be taken out of the regression and the model should be reestimated.

MARRIED is also positively significant. A married person or couple may have a larger household, and therefore needs a higher income. The variable *CITY* is also positively significant. People living in urban areas tend to face higher costs of living.

The regional variable *SOUTH* is not significant, but the variable *COLOR* is *negative* significant. This indicated that colored persons has a lower income, and maybe also less well-paid jobs.

Observe finally the coefficient of determination is quite low at 0.22. This is an issue frequently encountered in regressions with a high number of observations. However, the influence on the significance of the variables is limited.

3. Integrated Statistical Modeling and the use of Regression

Let us use our knowledge from the previous topics in the courses in Statistics I and II to perform an integrated model sequence. After having outlined for example a macroeconomic theory for example for the money demand, the consumption function or the investment function we now want to perform a statistical investigation. Let us assume that we have found some statistics from the national statistical bureau. Then such an analysis is split into two parts namely the ***descriptive statistical part*** and *a* ***regression part***. These parts could contain:

Descriptive statistical part

Here the following should be considered:

- Set up some nice time series or cross-section plots
- Compute *descriptive statistics* and comment on the evolution of the data
- If we use cross-section data: Draw Box-plot(s) and comment on data
- If we use time series data: Look for special events (like the 2007 recession) and consider the issue of seasonality

Regression part

- *Set up a matrix of correlation.* Identify the variables with the highest correlation to y and comment. Discuss signs and relate to the prior from economic theory. Do we find what we expect? As an alternative we could calculate the *variance inflation factor* for variable j defined as $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is coefficient of determination for the regression model that related x_j to all other independent variables x. If

multicollinarity is present the VIF_j will be high (R_j^2 near one) and vice versa. This measure is provided by Megastat.

- Could multicollinarity be present (correlation among *x* variables).
- Based on an initial estimation of the full model a model selection is undertaken. During this process we should observe:
    - We attempt to obtain the simplest model with the highest R^2 coefficient.
    - We attempt to minimize the "standard error of regression" shown in the Excel or Megastat output.
    - We attempt to eliminate multicollinarity.
    - All t-statistics should be significant (p-value < 0.05).
- For the final model some selected highlights of the model control can be shown.

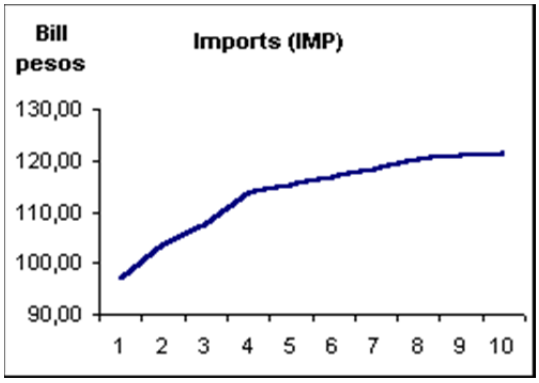
These things can all be undertaken by the use of Excel. Let us perform the regression part of this analysis on a small artificial data set. We want estimate an import *(IMP)* function for the artificial nation "Ruritania" for a 10 year period. We assume that the import depend on money supply *(MS)*, gross domestic product *(GDP)* the exchange rate of US dollar versus the local Peso *(USDvP)*, and finally the interest rate *(RENT)*. Below we find the statistics:

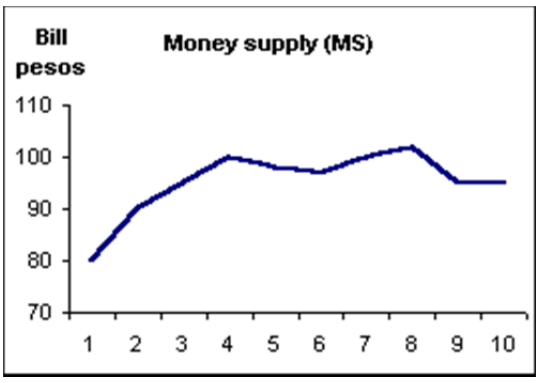
Data

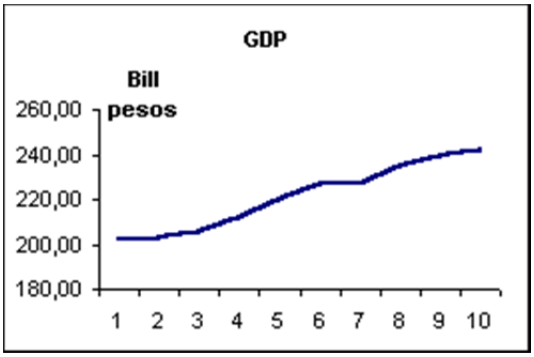
Year	*Imports (IMP) bill. pesos*	*Money supply (MS) bill. pesos*	*GDP* bill. pesos	*Exchange rate USD per Pesos*	*Interest rent (RENT)*
1	97.14	80	202.40	7.45	7.5
2	103.63	90	203.00	7.12	6.7
3	107.65	95	205.50	7.01	6.4
4	113.81	100	212.10	6.85	6.1
5	115.32	98	219.80	7.02	6.3
6	116.96	97	226.80	7.56	6.4
7	118.46	100	227.40	7.62	6.3
8	120.47	102	235.20	7.44	6.2
9	121.21	95	239.80	7.23	6.7
10	121.40	95	242.40	7.02	6.7

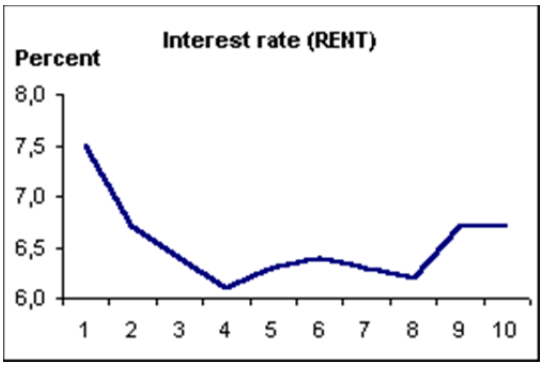
What should we expect from the theory of macroeconomics? When money supply increases, so do demand, so imports should increase. The same holds for GDP. If the interest rate decreases it will be cheaper to lent money. So a low interest rate should stimulate imports. Here we expect negative correlation. Correlation on exchange rate depends on the definition of the exchange rate. Here a low exchange rate should make imports cheaper. So here we expect a negative relation.

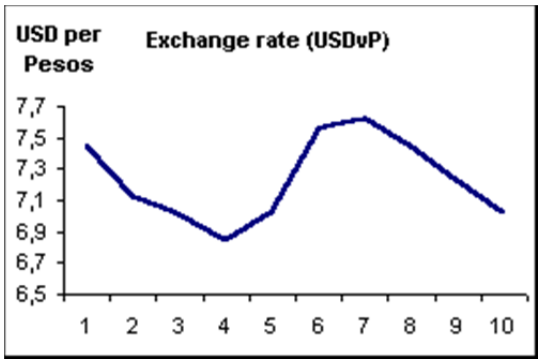
Let's look at some plots











From the plots we can observe that imports and GDP and money supply should be positively correlated. Further imports and the exchange rate as well as the interest rate should be negatively correlated. Notice that the exchange rate and the interest rate have a very similar pattern. If they are correlated with imports as well as with them self we observe a problem of *multicollinarity.* We want to estimate a model of the form:

$$IMP_t = \beta_0 + \beta_1 MS_t + \beta_2 GDP_t + \beta_3 USDrP_t + \beta_4 RENT_t + \varepsilon_t$$

Expected signs: (+) (+) (−) (−)

Let us first look at the matrix of correlation:

	IMP	*MS*	*GDP*	*USDvP*	*RENT*
y: Imports *(IMP)*	1.00				
x1: Money Supply *(MS)*	0.81	1.00			
x2: GDP	0.92	0.53	1.00		
x3: Exchange rate *(USDvP)*	0.06	−0.08	0.21	1.00	
x4: Interest rate *(RENT)*	−0.62	***−0.96***	−0.27	0.20	1.00

Notice, that many of our observations from the plots are confirmed. Besides from the exchange rate all variables are highly correlated with imports. Further, we were wrong with the correlation between the exchange rate and imports. We also observe a severe correlation between the money supply and the interest rate (−0.96). Also among GDP and money supply (0.53) multicollinarity is observed. Let us show the results from the estimation of the model:

Regression Statistics	
Multiple R	1.00
R-squared	0.99
Adjusted R-square	0.99
Standard Error	1.00
Observations	10

This is very high!

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	4	604.62	151.16	152.00	0.00
Residual	5	4.97	0.99		
Sum	9	609.60			

	Coefficient	*Standard deviation*	*t-stat*	*P-value*	*Lower 95 %*	*Upper 95 %*
Constant	78.87	112.72	0.70	0.52	-210.89	368.62
X1: MS	-0.02	0.73	-0.03	0.98	-1.90	1.86
X2: GDP	0.45	0.09	4.96	0.00	0.21	0.68
X3: USDvP	-1.12	1.50	-0.75	0.49	-4.98	2.73
X4: RENT	-8.26	10.26	-0.81	0.46	-34.63	18.11

We have in order to save space omitted the residuals diagrams. The coefficient of determination is very high and from the ANOVA table it is observed that the F-test is significant. Consequently it is meaningful to estimate the model.

However the model is very poor! The only variable that is significant is GDP. All other variables are not significant. The money supply even takes the wrong sign!

In order to proceed we will try to eliminate the most severe problem of multicollinarity namely among the money supply and the interest rate. So we estimate the model without the

interest rent. We exclude the interest rate because the money supply is higher correlated with the other variables than the interest rate.

We obtain the following output from Excel, and let us in this case include the residual analysis in the output.

Model without the Interest Rate

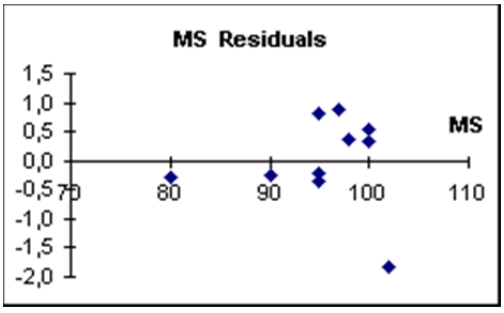
Regression Statistics	
Multiple R	1.00
R-squared	0.99
Adjusted R-square	0.99
Standard Error	0.97
Observations	10

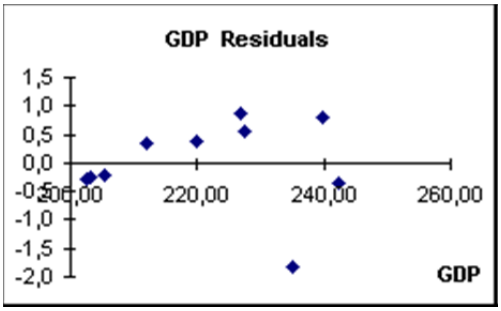
This is smaller than above

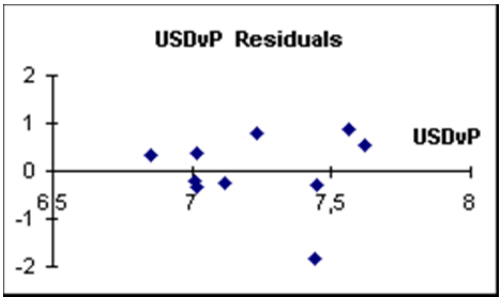
ANAVA

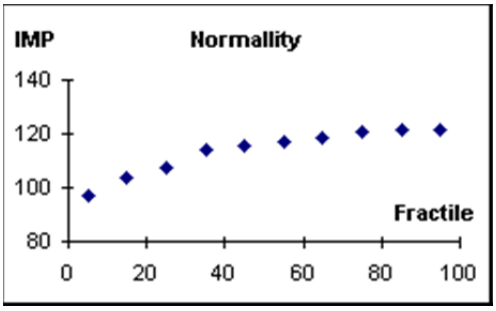
	df	*SS*	*MS*	*F-test*	*P-value*
Regression	3	603.98	201.33	215.04	0.00
Residual	6	5.62	0.94		
Sum	9	609.60			

	Coefficient	*Standard deviation*	*t-stat*	*P-value*	*Lower 95 %*	*Upper 95 %*
Constant	-11.52	10.33	-1.12	0.31	-36.80	13.75
X1: MS	0.57	0.06	9.18	0.00	0.42	0.72
X2: GDP	0.38	0.03	14.29	0.00	0.31	0.44
X3: USDvP	-1.72	1.27	-1.35	0.22	-4.82	1.38





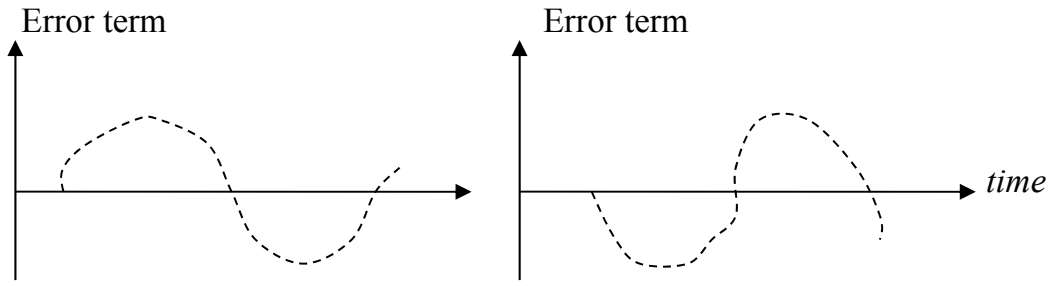




Compared to the initial estimation we now observe the correct sign for the money supply. Notice that the size of the coefficient of the GDP-variable has remained quite constant. Both variables are now significant. The sign on the exchange rate variable is as expected, but it is not significant. This means that the model should be reestimated without this variable. This will due to low correlation with MS and GDP as seen from the matrix of correlation not affect these variables. Finally the analysis of residuals as well as the plot for normality looks quite satisfactory.

4. Analyzing Autocorrelation and the Durbin-Watson Test

Autocorrelation occurs in non-stationary **time series**[1] where the variables are dependent in time. Autocorrelation may be either positive or negative of nature. Examples are given below:



Positive autocorrelation *Negative autocorrelation*

We test for autocorrelation be setting up the *Durbin-Watson* test. We calculate the tester:

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}$$

This expression is based on the estimation of the regression: $\varepsilon_t = p\varepsilon_{t-1} + v_t$ where the last term is ”the error term of the errors”. We can state the hypothesis as:

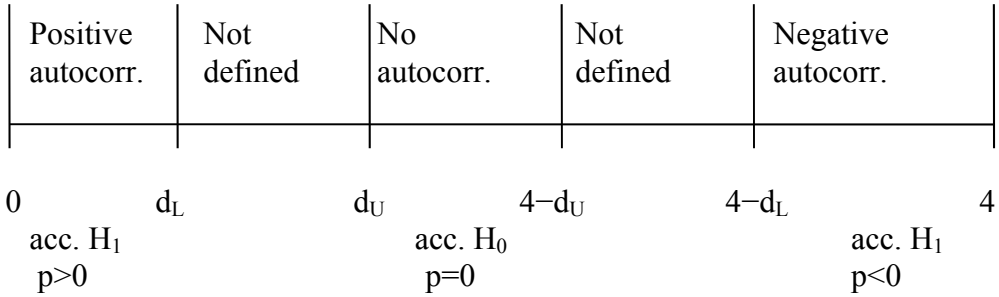
[1] Notice, that the test for autocorrelation only has a meaning, when we work with time series data, and NOT when we work with cross-section data. For example, in the latter case, data may be listed in alphabetic order. If we for example applied the test on regional statistics for Germany, performing the test would imply that Mainz and Munich would be directly related although the distance between the two cities is several hundred kilometers.

H_0: The error terms are not autocorrelated $(p = 0)$

H_1: The error terms are autocorrelated $(p \neq 0)$

The Durbin-Watson test is a two-sided test, where the alternative hypothesis (H_1) is not defined consistently. This is so because under H_0 the assumption to the error term is by itself not fulfilled. This is exactly what we want to test for!

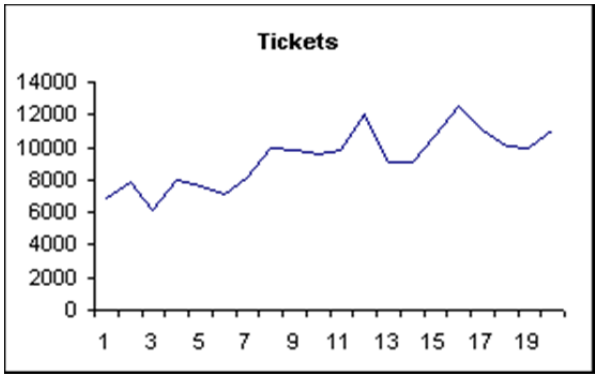
The distribution for the Durbin-Watson test is non-standard and found in Appendix II at the end of these notes. *k* is the number of explanatory variables (the number of X's). There are two critical values to be found named d_L and d_U. The range of the critical value is between 0 and 4. The interpretation can be summarized in the following figure:

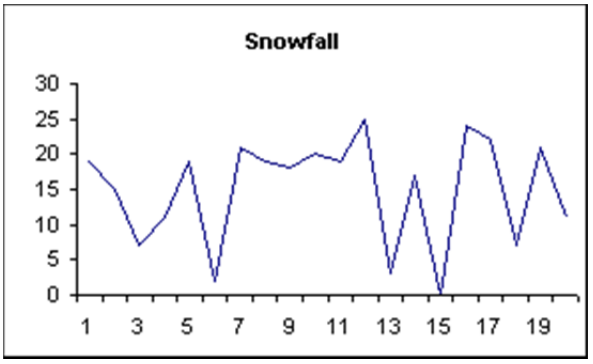


Example

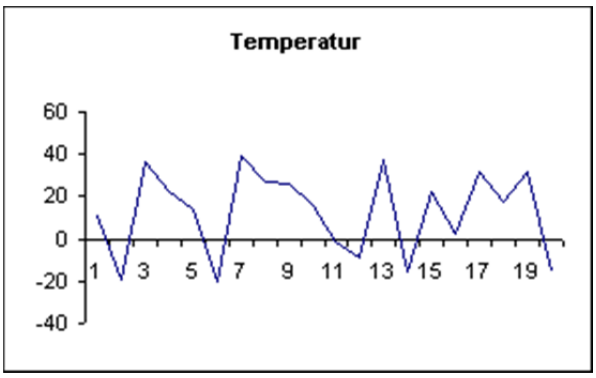
A ski resort wants to determine the effect that the weather have on the sales of lift tickets during the Christmas week. Weekly sales of ski lifts tickets *(y)* are assumed to depend on total snowfall in inches (x_1) and the average temperature in Fahrenheit (x_2). For a data set ranging over 20 years we obtain:

Tickets (y)	Snowfall (x_1)	Temperature (x_2)
6835	19	11
7870	15	-19
6173	7	36
7979	11	22
7639	19	14
7167	2	-20
8094	21	39
9903	19	27
9788	18	26
9557	20	16
9784	19	-1
12075	25	-9
9128	3	37
9047	17	-15
10631	0	22
12563	24	2





11012	22	32
10041	7	18
9929	21	32
11091	11	-15



From the plots it is observed that the relation among the variables not is optimal. So we do not expect the most significant result. This is also confirmed by the matrix of correlation shown below. We obtain this from Excel by use of the command: ***Data/data analysis/correlation***:

	Tickets	*Snowfall*	*Temperature*
Y: Tickets	1.00		
X1: Snowfall	0.33	1.00	
X2: Temperature	-0.11	-0.02	1.00

Let us now estimate a model of the form:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \qquad \text{where } t = 1,2,\ldots,20$$

From Excel we obtain:

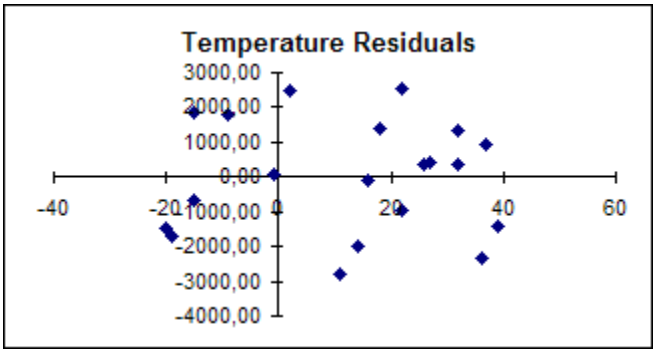
Regression Statistics	
Multiple R	0.35
R square	0.12
Adjusted R square	0.02
Standard Error	1711.68
Observations	20

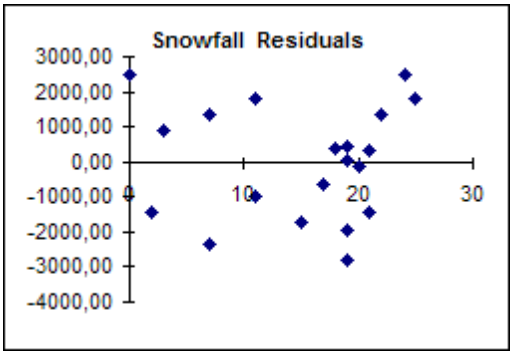
This is a poor correlation

ANOVA

	df	***SS***	***MS***	***F-test***	***P-value***
Regression	2	6793798.5	3396899.1	1.16	0.34
Residual	17	49807214	2929836.1		
Sum	19	56601012.2			

	Coefficient	***Standard Deviation***	***t-stat***	***P-value***	***Lower 95%***	***Upper 95%***
Constant	8308.01	903.73	9.19	0.00	6401.31	10214.71
X1: Snowfall	74.59	51.57	1.45	0.17	-34.22	183.41
X2: Temperature	-8.75	19.70	-0.44	0.66	-50.33	32.82





This is not a very good result, and much worse than our expectations from the plots! The F-test is not significant ($p = 0.34 > 0.10$), so the overall model is not significant. Further, only the constant term is significant. It looks like that neither snowfall nor temperature has an influence on the sales of tickets.

The plots of residuals are also not very nice! Both plots reveal some kind of systematic behavior. Let us perform the Durbin-Watson test first by calculation of the formula given above in Excel. We then obtain:

ε_t	ε_{t-1}	$(\varepsilon_t - \varepsilon_{t-1})^2$	$(\varepsilon_t)^2$
-2793.99			7806391.51
-1723.23	-2793.99	1146528.83	2969525.68
-2342.03	-1723.23	382911.49	5485102.65
-956.95	-2342.03	1918431.85	915762.73
-1963.73	-956.95	1013597.71	3856238.75
-1465.27	-1963.73	248460.53	2147024.00
-1439.07	-1465.27	686.38	2070933.63
414.07	-1439.07	3434133.96	171452.12
364.91	414.07	2416.75	133157.32
-102.82	364.91	218765.62	10571.25
49.96	-102.82	23341.64	2496.31
1823.37	49.96	3144985.16	3324691.68
920.10	1823.37	815908.57	846578.74
-660.40	920.10	2497979.80	436131.76
2515.57	-660.40	10086807.91	6328096.53
2482.26	2515.57	1109.74	6161605.15
1343.06	2482.26	1297779.75	1803801.32
1368.40	1343.06	642.44	1872527.09
334.65	1368.40	1068645.60	111990.59
1831.16	334.65	2239532.65	3353135.13
	Sum	29542666.38	49807213.95

$$DW = \frac{29542666.38}{49807213.95} = 0.59$$

It is assumed that $n=20$ and $k=2$.
With a level of significance equal to 0.05, we have from the critical values in Appendix II that

$D_L = 1.10$ $D_U = 1.54$

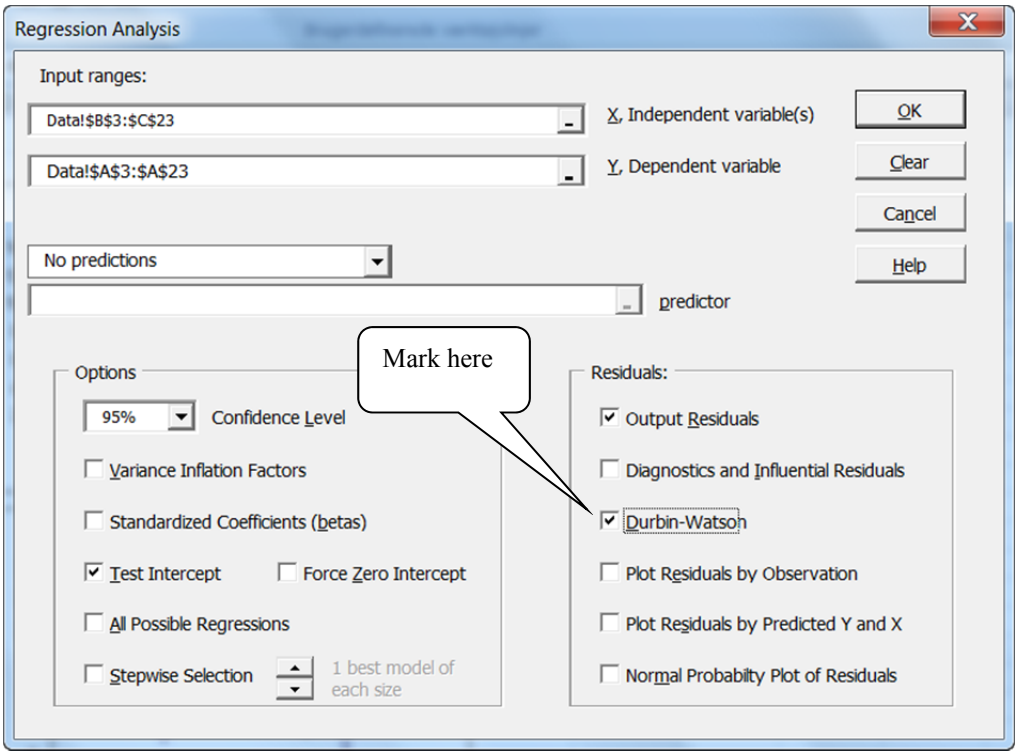
Hypothesis:

H_0: No first order autocorrelation

H_1: Positive first order autocorrelation

As DW < D_L H_1 is accepted.

Alternatively, we can find the DW-value by use of Megastat. Here the test is much easier to perform. In the menu for *regression,* a label with the text *Durbin-Watson* can be found. Just mark the label, and the test will be performed. The menu looks as:



We then find that positive autocorrelation is present. How do we solve the problem? A solution could be to include a ***positive linear trend***. This is a variable taking the values $T = 1,2,..,20$. It is a strongly positive variable. We must from the plots expect this variable to be strongly correlated with the sales of tickets as this has a positive trend.

The result with inclusion of a positive linear trend from Excel is:

Regression Statistics		
Multiple R	0.86	This is has increased a lot!
R square	0.74	
Adjusted R square	0.69	
Standard Error	957.24	This has decreased a lot!
Observations	20	

ANOVA

	df	***SS***	***MS***	***F-test***	***P-value***	
Regression	3	41940217.4	13980072.5	15.26	0.00	Significant!
Residual	16	14660794.8	916299.676			
Sum	19	56601012.2				

	Coefficient	***Standard deviation***	***t-stat***	***P-value***	***Lower 95 %***	***Upper 95 %***
Constant	5965.59	631.25	9.45	0.00	4627.39	7303.78
X1: Snowfall	70.18	28.85	2.43	0.03	9.02	131.35
X2: Temperature	-9.23	11.02	-0.84	0.41	-32.59	14.13
X3: Trend	229.97	37.13	6.19	0.00	151.25	308.69

Compared to the first regression an improvement can be observed. Snowfall is now significant, but temperature has no effect on the model, and should be excluded. Further the coefficient of determination has increased and the standard error has decreased. Consequently this is the model to be preferred.

For this model the Durbin-Watson test can also be undertaken. This will result in a DW-value equal to 1.88. Now $k=3$ because the trend is included. The critical values can again be found be use for the appendix. In this case $d_L=0.998$ and $d_U=1.676$. As $1.676<1.88$ no autocorrelation is observed. The inclusion of the trend variable eliminates the presence of autocorrelation.

Appendix I: Estimating the Multiple Regression Model*

The multiple regression model with *k* regressors can in a mathematical way be stated as:

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

We consider the same assumptions as with the simple regression analysis with regard to the error term. As stressed above it is difficult to solve for the *β's*. This is so because we now consider a system of *(k+1)* normal equations. So the regression model is a polynomial of degree *(k+1)*. In the case where *k=2* (2 independent *x* variables) it is possible to show how the multiple regression plan is set up.

Figure 1 on page 17 undertakes this task and compare with the simple linear regression.

In order to solve the system we use matrix algebra. We use a capital boldface letter to denote a matrix such as **X,** and boldface small letter such as **y** to denote a vector. Let us return to the example above and add an extra variable x_2 = [70; 65; 60; 75; 50; 45; 40] to the data set. Then on matrix form:

$$\mathbf{X} = \begin{bmatrix} 1 & 100 & 70 \\ 1 & 200 & 65 \\ 1 & 300 & 60 \\ 1 & 400 & 75 \\ 1 & 500 & 50 \\ 1 & 600 & 45 \\ 1 & 700 & 40 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{bmatrix}$$

Notice that the dimension of the matrix is determined by the number of rows and columns in the matrix. The matrix **X** has a number of rows equal to the number of observations *n* and columns equal to the number of *x* variables *k* plus 1 (this is due to the constant term). So the dimension is equal to $n \times (k+1)$. Similarly for the vector **y**. On matrix form our example model is:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad \text{where } \boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2] \text{ dimension } 1 \times (k+1)$$

and **ε** is a vector with dimension similar to **y**.

To use **X** and **y** to calculate **β** we have to define the *transpose matrix* of **X** denoted by **X'**. The *transpose* of a matrix is formed by interchanging the rows and columns of the matrix. Consequently the dimension of the transpose is the reverse. Let us now solve the expression

disregarding $\boldsymbol{\varepsilon}$ for $\boldsymbol{\beta}$, and multiply by the transpose matrix in order to have match between the dimension of rows and columns. Then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \quad \Leftrightarrow \quad \boxed{\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

Technically we have minimized the squared sum of the errors or

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{or} \quad (\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})$$

The minimum occurs where the partial derivatives with respect to $\boldsymbol{\beta}$ are zero. Doing this the vector of partial derivatives is

$$\mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Solving for $\boldsymbol{\beta}$ the expression in the box above is obtained. Undertaking this manually is especially when the number of $k > 2$ and $n > 20$ is extremely complicated, so luckily we have computers to undertake the problem! Finally, we can define

Explained variation: $SSR = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$

Unexplained variation: $SSE = \sum y_i^2 - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$

Finally let us look at the assumptions for the multiple regression model with regard to the error term:

- The expected value of the residuals: $E(\mathbf{e})$
- The covariance of matrix of $\mathbf{e}$ is: $cov(\mathbf{e}) \equiv E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}$

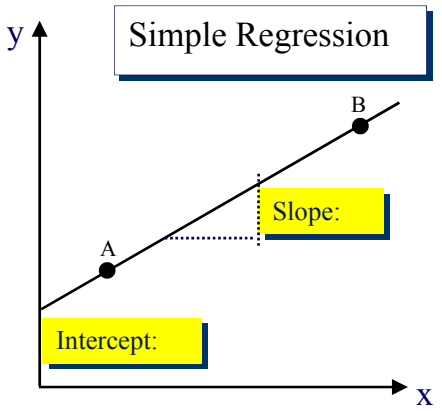
Where $\mathbf{I}$ is the identity matrix. For example a 3×3 identity matrix looks like

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

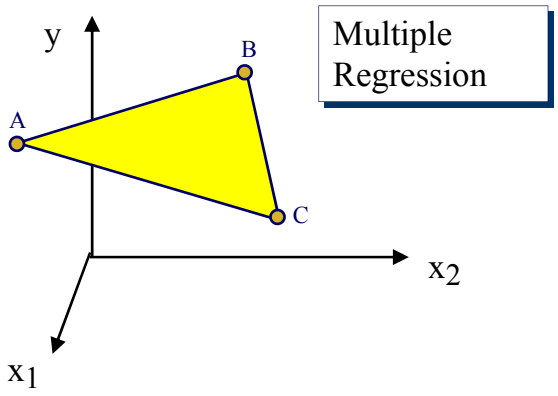
So the interpretation of the covariance statement is just that the observations are independent.

The expected value of y or $\hat{y}$ is as a consequence equal to: $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

Figure 1: Illustration of the simple regression (line) and the multiple regression (plane)



Any two points (A and B), or an intercept and slope (β_0 and β_1), define a *line* on a two-dimensional surface.



Any three points (A, B, and C), or an intercept and coefficients of x_1 and x_2 (β_0 , β_1 , and β_2), define a *plane* in a three-dimensional surface.

Appendix II: Critical Points for the Durbin-Watson Test Statistic

95 % level ($\alpha = 0.05$)

n	k = 1		k = 2		k = 3		k = 4		k = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.172	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820

Source:

Durbin, J. and G. S. Watson, 1951, *Testing for Serial Correlation in Least Squares Regression,* Biometrika 30, pp. 158−178.